Exploring new methodologies to identify disease-associated variants in African populations through the integration of patient genotype data and clinical phenotypes derived from routine health data: A case study for Type 2 Diabetes Mellitus in patients in the Western Cape Province, South Africa

By

Tsaone Tamuhla

TMHTSA001

Thesis presented for the degree of
DOCTOR OF PHILOSOPHY
in the
Department of Integrative Biomedical Sciences
at the
Faculty of Health Sciences
University of Cape Town

Submission date: 13th February 2023

Primary supervisor: Professor Nicola C Tiffin
Co-supervisor: Professor Nicola J Mulder

Declaration

This thesis is presented in fulfilment of the requirements for the degree of Doctor of Philosophy (PhD) in the Department of Integrative Biomedical Sciences at the Faculty of Health Sciences, University of Cape Town. The work included in this thesis is original research, and has not, in whole or in part, been submitted for another degree at this or any other university. The contents of this thesis are entirely the work of the candidate or, in the case of multi-authored published papers, constitutes work for which the candidate was the lead author.

This thesis includes published manuscripts, as per general provision 6.7 in the General Rules for the Degree of Doctor of Philosophy (PhD) of the University of Cape Town. I confirm that I have been granted permission by the University of Cape Town's Doctoral Degrees Board to include the following publications in my PhD thesis, and where co-authorships are involved, my co-authors have agreed that I may include these publications. The following manuscripts (all published) are included in the thesis, and are presented as self-contained chapters in the following order:

1. An e-consent framework for tiered informed consent for human genomic research in the global south, implemented as a REDCap template. BMC Med Ethics. 2022;23: 119. doi:10.1186/s12910-022-00860-2 (chapter 2).
2. Diabetes in a TB and HIV-endemic South African population: Analysis of a virtual cohort using routine health data. PLOS ONE. 2021;16: e0251303. doi:10.1371/journal.pone.0251303 (chapter 3).
3. Risk factors for COVID-19 hospitalisation and death in people living with diabetes: A virtual cohort study from the Western Cape Province, South Africa. Diabetes Res Clin Pract. 2021;177: 108925. doi:10.1016/j.diabres.2021.108925 (chapter 4).

Acknowledgements

To my supervisor Prof Nicki Tiffin, thank you responding to my email and setting me on this amazing trajectory. Words are not enough to express my sincere gratitude for the amazing opportunities you have afforded me and above all else thank you for being a great mentor, you made my PhD an amazing experience. (Word of the day is amazing)

To Prof's Joel Dave and Peter Raubenheimer, thank you for your invaluable contribution to my PhD and a special thank you to Dr Maleeka Abrams for tirelessly assisting with collecting the buccal swab samples.

To my co-supervisor and Head of Division Prof Nicola Mulder, thank you for all your support during my time at CBio.

I would also like to thank H3ABioNet for generously funding my PhD.

I would also like to thank all the participants who volunteered to be in my study. Without you, this project would not have been possible.

Last but not least, I would like to thank my family and friends for supporting me tirelessly throughout this journey.

To my mother Neo and my sisters Ludo and Isabella, we don't choose our family, but if we could, I would choose you over and over. I am forever indebted to you for selflessly filling in the gaps of my absence with Rori and in doing so allowing me freedom to pursue my dreams.

To my Rorisang, I dedicate this work to you. May you be inspired to live your life purposefully, always striving to be the very best that you can be.

“A dream delayed is not a dream denied” ~ Annonymous

Abstract

There is poor knowledge on the genetic drivers of disease in African populations and this is largely driven by the limited data for human genomes from sub-Saharan Africa. While the costs of generating human genomic data have gone down significantly, they are still a barrier to generating large scale African genomic data. This project is therefore a proof-of-concept pilot study that demonstrates the implementation of a cost-effective, scalable genotyped virtual cohort that can address population level genomic questions.

We optimised a tiered informed consent process that is suitable for the cohort study design and adapted it to conducting human genomic research in the African context. We used an existing dataset to explore statistical methods for modelling longitudinal routine health data into a standardised phenotype for genome wide association studies (GWAS). We then conducted a feasibility study and piloted the tiered informed consent process, DNA collection by buccal swab and DNA extraction from buccal swabs and peripheral blood samples. DNA samples were genotyped for approximately 2.2 million variants on the Infinium™ H3Africa Consortium Array V2. Genotyping quality control (QC) was done in Plink 1.9 and genome wide imputation on the Sanger Imputation Service. We demonstrated successful variant calling and provide aggregate statistics for known aetiological variants for type 2 diabetes and severe COVID-19 as well as demonstrating the feasibility of running nested case-control GWAS with these data.

We demonstrate the use of routine health data to provide complex phenotypes to link to genotype data for both non-communicable diseases (diabetes) and infectious diseases (Tuberculosis, HIV and COVID-19). 459 participants consented to providing a DNA sample and access to their routine health data and were included in the feasibility study. A total of 343 DNA samples and 1782023 genotyped variants passed quality control and were available for further analysis. While most of the cohort population clustered with the 1000 genomes African population, principal component analysis showed extensive population admixture. For the COVID-19 analysis, we identified 63 cases of severe COVID-19 and 280 controls, and for the type 2 diabetes analysis we identified 93 cases and 250 controls using the routine health data of participants in the cohort. While the sample sizes were insufficient for a GWAS we were able to evaluate known type 2 diabetes mellitus and COVID-19 variants in the study population.

We have described how we conceptualised and implemented a genotyped virtual population cohort in a resource constrained environment, and we are confident that this design and implementation are appropriate to scale up the cohort to a size where novel health discoveries can be made through nested case-control studies. In the interim we demonstrate the analysis and validation of aetiological variants identified in other studies and populations.

List of abbreviations

ACE-I	Angiotensin converting enzyme inhibitor
AIDS	Acquired Immune Deficiency Syndrome
ARB	Angiotensin receptor blocker
ART	Antiretroviral treatment
BMI	Body mass index
CKD	Chronic Kidney disease
CMA	Continuous medication availability
COPD	Chronic obstructive pulmonary disease
COVID-19	Corona virus disease 2019
CVD	Cardiovascular disease
DM	Diabetes Mellitus
DNA	Deoxyribonucleic acid
e-consent	Electronic capture of consent process
GWAS	Genome wide association study
H3Africa	Human Hereditary Health in Africa
HbA1c	Glycated haemoglobin A1c
HIV	Human Immunodeficiency Virus
HPT	Hypertension
ICU	Intensive care unit
ID	Identification
IRB	Institutional review board
LD	Linkage disequilibrium
MAF	Minor allele frequency
MDS	multidimensional scaling
MPR	Medication possession ratio
NCD	non-communicable disease
NGSP	National Glycohemoglobin Standardization Program
NNRTI	Non-nucleoside reverse transcriptase inhibitors
OR	Odds ratio
PCA	Principal component analysis
PCR	Polymerase chain reaction
PDC	Proportion of days covered
PDF	Portable document format
PHDC	Provincial health data center

PI	Protease inhibitors
PID	Unique record name
PLWD	People living with diabetes
PLWH	People living with HIV
PMI	Patient master index
QC	Quality control
REDCap	Research Electronic Data Capture
SNP	Single nucleotide polymorphism
SSA	Sub-Saharan Africa
T2DM	Type 2 Diabetes Mellitus
TB	Tuberculosis
WCGH	Western Cape Government Health

Table of Contents

List of figures

List of tables

1. Literature review

1.1. Introduction

1.1.1. Diabetes in sub-Saharan Africa

Sub-Saharan Africa (SSA) is currently faced with a dual burden of infectious and increasing prevalence of non-communicable diseases (NCD's) (Birungi et al., 2021; Garrib et al., 2019) and it is predicted that by 2030, NCD's will surpass infections as the leading cause of morbidity and mortality in the region (Bigna & Noubiap, 2019; Gouda et al., 2019). Notably, cardiovascular disease and diabetes associated morbidity and mortality are putting a strain on already struggling public health systems (Adeniyi, Yogeswaran, Longo-Mbenza, & Goon, 2016; Jaffar et al., 2021). An estimated 24 million African adults (20 – 79 years) are currently living with diabetes and with the number projected to reach 55 million by 2045 (*IDF Diabetes Atlas 10th Edition*, n.d.), the region is facing an impending diabetes epidemic (Gill et al., 2008; Levitt, 2008).

More than 90% of diabetes in SSA is type 2 diabetes mellitus (T2DM) (Hall et al., 2011) which is thought to be largely fuelled by lifestyle changes brought about by a surge in rural urban migration (Bertram et al., 2013) and an aging population (Pastakia et al., 2017). The situation is further exacerbated by the interaction of diabetes with infectious disease including HIV (Bam et al., 2020; Birungi et al., 2021; Bosire, 2021; Levitt et al., 2016), Tuberculosis (Al-Rifai et al., 2017; Berkowitz et al., 2018; Gennaro et al., 2019) and more recently COVID-19 (Apicella et al., 2020; Bramante et al., 2021; McGurnaghan et al., 2021) which further complicate the pathogenesis of T2DM in SSA.

While it is well established that T2DM is a caused by a combination of lifestyle and genetic factors (Bertram et al., 2013; Gill et al., 2008; Hall et al., 2011; Levitt, 2008), most African studies have focused on the lifestyle drivers of the disease (Adeniyi, Yogeswaran, Longo-Mbenza, & Goon, 2016; Adeniyi, Yogeswaran, Longo-Mbenza, Goon, et al., 2016; Amberbir et al., 2019; Manyema et al., 2015) and there remains a dearth of knowledge on the genetic drivers of T2DM in Africans.

1.1.2. Genetics of T2DM

Genome wide association studies (GWAS) have been used widely to study the genetics of T2DM in different populations. This approach has enjoyed wide success as there are currently over 700 known risk loci for T2DM identified mainly in European and Asian populations (DeForest & Majithia, 2022; Irgam et al., 2021; Ishigaki et al., 2020; Spracklen et al., 2017) with the former dominating. While these studies have helped to advance our current knowledge on the genetics of T2DM, they are limited in their lack of representation from other populations, most notably continental Africans (DeForest & Majithia, 2022). In their analysis, Deforest & Mijita (2022) identified only one study (J. Chen et al., 2021) which included data from SSA populations in their meta-analysis (DeForest & Majithia, 2022). However, the lack of African populations in multi-ancestry meta-analyses is largely due to the unavailability of African data.

To date, most of the genetic research on T2DM in Africans have been candidate gene studies done in North African populations (Yako et al., 2016) and the few done in populations from SSA have been dominated by participants from Nigeria, Ghana, Kenya and South Africa (Adeyemo et al., 2019; G. Chen et al., 2007; J. Chen et al., 2019, 2021; Chikowore et al., 2022; Rotimi et al., 2004). In addition, only two adequately powered T2DM GWAS (Adeyemo et al., 2019; J. Chen et al., 2019) have been performed in African populations. In their T2DM GWAS on 5321 participants from Nigeria, Kenya and Ghana, Adeyemo and colleagues (2019) identified the novel African specific T2DM locus ZRNAB3 (Adeyemo et al., 2019). Similarly, Chen and colleagues (J. Chen et al., 2019) identified novel (rs73284431) in their GWAS of 4349 individuals from South Africa, Nigeria, and Kenya. While these studies have added significantly to the body of knowledge, the limited diversity in the sub-Saharan populations being studied is worrying in a region that boasts the greatest genetic diversity out of all populations (Choudhury et al., 2018).

1.1.3 Genetic diversity in African populations

The genetic diversity of African populations extends beyond inter-continental comparisons as significant differences also exist between African ancestry populations (Choudhury et al., 2017, 2021; Kamiza et al., 2022; Patin et al., 2017). Numerous studies on the population structure in different African populations have revealed highly heterogenous populations with extensive admixture particularly in Southern Africa (Chimusa et al., 2013; Choudhury et al., 2018; Daya et al., 2013; de Wit et al., 2010; Petersen et al., 2013). This genetic diversity has implications for both GWAS and post-GWAS analysis in African populations (Teo et al., 2010). This is because the current GWAS methods are based on the premise of a genetic homogeneity in the population being studied and applying them indiscriminately to heterogenous population could result in loss of valid genetic associations (Kulminski et al., 2016; Teo et al., 2010). In addition, this assumption of homogeneity, means that results from one populations can be inferred to another. However, recent studies have shown that there is poor transferability of genetic risk scores (GRS) and polygenic risk scores (PRS) from European ancestry populations to African populations (Chikowore et al., 2022; Kamiza et al., 2022). Additionally, in their investigation, Kamiza and colleagues (2022) showed that even between African populations, there is limited transferability of GRS predictions (Kamiza et al., 2022) further highlighting the urgent need for large scale generation of African genomic data with a wider population coverage (Fatumo, 2020).

1.1.4. Barriers to doing human genomic research in Africa

Given the significant potential for African genomic data to not only address African specific problems, but to also close gaps in global health issues (Fatumo, 2020; Gurdasani et al., 2015; N. Mulder, 2017; Ramsay, 2012) efforts have been made to identify and address the barriers to doing genomic research in African populations (Adebamowo et al., 2018; *Policy Paper: A Framework for the Implementation of Genomic Medicine for Public Health in Africa | The AAS*, n.d.; Wonkam, 2021).

Generating genomic data is a resource intensive process. One of the most cited barriers to doing human genomic research in Africa, is the general lack of resources needed to run the projects, the biggest being the set up and maintenance of adequate infrastructure (Adebamowo et al., 2018; Jongeneel et al., 2022; *Policy Paper: A Framework for the Implementation of Genomic Medicine for Public Health in Africa | The AAS*, n.d.; Ramsay, 2012). While there are various infrastructure needs, the biggest pertains to the long-term storage of participant samples. This is usually done in biobanks such as the UK Biobank (Sudlow et al., 2015) and Biobank Japan (Nagai et al., 2017) which have the capacity to store an array of both baseline and follow-up biospecimen. These biobanks have worked to advance human genomic research in their populations as they enable the establishment of large disease agnostic population cohorts through which multiple diseases can be studied from the same resources (Ishigaki et al., 2020).

Apart from the cost associated with setting up and maintaining biobanks, there are ethical considerations to be made when setting up such infrastructure, especially in countries that do not have strong laws governing the protection of personal information (Christoffels & Abayomi, 2020; de Vries et al., 2017; Grady et al., 2015; Mikkelsen et al., 2019). To avoid mis-use and potential harm to the participants, there needs to be adequate governance structures including informed consent in place to ensure that participant samples are only used for the purposes the participant consented to (de Vries et al., 2017; Nembaware et al., 2019; Tiffin, 2018)

The inadequate sample sizes in African genomic research are not always due to lack of funding. Some researchers fail to enrol adequate numbers into their studies and apart from research fatigue in some participants, inadequate participant recruit has also been attributed to lack of appropriate informed consent tools (Adebamowo et al., 2018). To avoid excluding qualifying participants, it has been suggested that consent forms be translated into the local language, so that language is not a barrier. However, this can be a challenging and costly process to implement particularly in settings where multiple languages are spoken (Adebamowo et al., 2018). At the same time there have been calls to move away from the widely used broad consent in genomic research (Grady et al., 2015; Maloy & Bass, 2020) and instead, use the tiered informed consent model which gives participants the autonomy to decide the secondary use of their data (Nembaware et al., 2019; Tiffin, 2018). In addition, because using participant centered informed consent models involves the patient in the decision making process regarding their data, it could also work to help build trust between participants and researchers and potentially improve their willingness to volunteer for genomic research activities (Adebamowo et al., 2018).

Genomic studies require robust health data which can be used to generate phenotypes for the genotype data. Data collection is an expensive undertaking especially in studies that require follow-up visits. To mitigate this, large established population cohorts have opted for the secondary use of electronic medical records to generate phenotype data for their cohorts (Brumpton et al., 2022). This is a cost-effective way to collect longitudinal phenotype data because follow-up data can be easily accessed from the updated electronic health records of consenting participants. While there have been significant improvements in the generation of electronic routine health data in African countries (Boulle et al., 2019; Todd & Mahande, 2020; Wabiri et al., 2019) in some countries it is still not adequate for secondary use in research. In such situations researchers often have to set up parallel data collection environments. This set-up is not only costly to the researcher, but it also burdens the participant who needs to give their data to multiple stakeholders.

In addition to phenotype data, researchers also need to generate genomic data. While the costs of generating genotype data have reduced significantly, they are still a significant barrier to generating large scale African genomic data. In addition, to the actual genotyping or sequencing costs, because most African countries lack genotyping and sequencing facilities, as such researchers incur additional shipping costs, which can be prohibitive especially when shipping samples outside the country (Croxton et al., 2017). In addition to the cost, shipping samples can also potentially affect their integrity especially if the proper shipping conditions are not observed and this works to further increase the cost through re-testing or re-shipping of samples (Croxton et al., 2017). There is therefore an urgent need to scale up local sequencing and genotyping capacities as this will drive down costs and enable more African researchers to conduct large scale genomic studies.

Genomic research has been a largely neglected area in African countries and it is not unexpected that there would be a shortage of adequately trained people to manage the various areas of the genomic research process (Adebamowo et al., 2018). In addition, the exodus of the few skilled individuals to countries in the global north has also been cited as a significant barrier to carrying out genomic research in African countries. To address some of these barriers and increase the uptake of genomic research in African populations, consortia such as the Human Hereditary and Health in Africa (H3Africa) have made significant contributions to advancing human genomic research in African countries (N. Mulder et al., 2018; N. J. Mulder et al., 2017). In addition, to enabling the establishment of African led genomic research in 30 countries across the continent, they have made significant contributions to developing cost-effective tools that are suitable for genomic research in African populations. The most notable is the Infinium™H3Africa Consortium Array (H3Africa) which is a custom micro array genotyping chip that is enriched for novel African specific variants (N. Mulder, 2017). In addition, they have also established an imputation service that uses an African specific reference panel to allow for optimal variant calling of informative African genotypes (Baichoo et al., 2018).

1.2. Study rationale

The dearth of knowledge of the genetic drivers of disease in African populations is largely driven by the limited data for human genomes from sub-Saharan Africa as this was a previously neglected area of research. This is because while the costs of generating human genomic data have gone down significantly in recent years, they are still a barrier to generating large scale African genomic data that can address population level questions in resource limited settings.

Single nucleotide polymorphism (SNP) genotyping is a cost-effective way of generating large scale genomic data and the recent availability of the Infinium™ H3Africa Consortium Array V2 (H3Africa chip), is making it possible to generate informative genotype data for genome wide association studies (GWAS) in African genomes. Most GWAS use the binary outcome of disease or no disease, however, in real life, disease phenotypes are more complex. This is because they are influenced by factors such as the environment, epigenome, and patient demographics. Therefore, using complex disease phenotypes that include these factors could potentially increase the utility of GWAS.

Routine health data collected through the course of chronic disease including diabetes, are a rich source of data that could be used to describe complex disease outcomes. This is because this data contains patient demographics, laboratory results, prescribed medications and hospital encounters for a patient over time. This longitudinal data therefore has greater utility than cross sectional data because it allows for the identification of time dependent disease patterns within an individual and allows a comparison to be made between individuals in a population.

1.3. Aim

The overarching aim of this research was to determine whether complex clinical phenotypes generated from routine health data can be used as a phenotype for genotype data in genome wide association studies in an African context.

1.4. Objectives

This PhD work was therefore a proof-of-concept study that aimed to demonstrate the creation of a virtual genotype cohort through the integration of complex clinical phenotypes from longitudinal routine health data and genotype data from the H3Africa chip using type 2 diabetes (an NCD) and COVID-19 (an infectious disease) as case studies.

This study had three main objectives:

Objective 1

To optimise and adopt a tiered informed consent that is suitable for the cohort study design and for conducting human genomic research in the African context.

- Describe and apply a participant-centred consent strategy for collecting and accessing sensitive genomic and health data.
- Recruit 300 participants who consent to the access and use of their individual longitudinal health data from the Provincial Health Data Centre and its linkage to their genotype data.
- Collect samples (two buccal swabs) from each participant and send for DNA preparation and storage.

Objective 2

To use an existing routine health data set to generate descriptive statistics from the demographic data and to describe diabetes profiles in the population from which sampling for the cohort will be done. Additionally, this data will also be used to explore statistical methods for modelling longitudinal data so that it can be used as a GWAS phenotype.

- Generate summary statistics from longitudinal clinical data for participants with T2DM in the Western Cape.
- Explore and describe statistical modelling methods for analysing patient outcome data using longitudinal clinical data.

- ❖ Apply an appropriate statistical model to the longitudinal clinical data from the T2DM patient data to generate patient profiles (disease phenotypes).

Objective 3

To demonstrate data integration by linking genotype data to routine health data to understand genetic and clinical drivers of health outcomes in an African virtual cohort.

- ❖ To generate genotype data for consenting participants using the H3Africa chip.
- ❖ Optimise and apply existing genotyping quality control and imputation methods and tools that are suitable for genotype data from the H3Africa chip.
- ❖ Demonstrate the feasibility of running nested case control GWAS with these data using T2DM and severe COVID-19 as phenotypes.

1.5. Structure of the thesis

This thesis comprises seven chapters. The first is an introductory chapter which consists of a literature review, the study rationale, aim and objectives. Chapters 2 to 6 address the different research objectives and chapter 7 is a general discussion. Chapters 2, 3 and 4 are published manuscripts. The candidate conducted all the data analyses, wrote the first drafts, addressed comments from reviewers and was first author for all the included manuscripts. The relevance of each manuscript to the thesis and more detailed author contributions are included at the start of each manuscript chapter.

References

Adebamowo, S. N., Francis, V., Tambo, E., Diallo, S. H., Landouré, G., Nembaware, V., Dareng, E., Muhamed, B., Odutola, M., Akeredolu, T., Nerima, B., Ozumba, P. J., Mbhele, S., Ghanash, A., Wachinou, A. P., & Ngomi, N. (2018). Implementation of genomics research in Africa: Challenges and recommendations. *Global Health Action*, *11*(1), 1419033. https://doi.org/10.1080/16549716.2017.1419033

Adeniyi, O. V., Yogeswaran, P., Longo-Mbenza, B., & Goon, D. T. (2016). Uncontrolled Hypertension and Its Determinants in Patients with Concomitant Type 2 Diabetes Mellitus (T2DM) in Rural South Africa. *PLOS ONE*, *11*(3), e0150033. https://doi.org/10.1371/journal.pone.0150033

Adeniyi, O. V., Yogeswaran, P., Longo-Mbenza, B., Goon, D. T., & Ajayi, A. I. (2016). Cross-sectional study of patients with type 2 diabetes in OR Tambo district, South Africa. *BMJ Open*, *6*(7), e010875. https://doi.org/10.1136/bmjopen-2015-010875

Adeyemo, A. A., Zaghloul, N. A., Chen, G., Doumatey, A. P., Leitch, C. C., Hostelley, T. L., Nesmith, J. E., Zhou, J., Bentley, A. R., Shriner, D., Fasanmade, O., Okafor, G., Eghan, B., Agyenim-Boateng, K., Chandrasekharappa, S., Adeleye, J., Balogun, W., Owusu, S., Amoah, A., … Rotimi, C. N. (2019). ZRANB3 is an African-specific type 2 diabetes locus associated with beta-cell mass and insulin response. *Nature Communications*, *10*(1), Article 1. https://doi.org/10.1038/s41467-019-10967-7

African region tops world in undiagnosed diabetes: WHO analysis. (2023, February 8). WHO | Regional Office for Africa. https://www.afro.who.int/news/african-region-tops-world-undiagnosed-diabetes-who-analysis

Al-Rifai, R. H., Pearson, F., Critchley, J. A., & Abu-Raddad, L. J. (2017). Association between diabetes mellitus and active tuberculosis: A systematic review and meta-analysis. *PloS One*, *12*(11), e0187967. https://doi.org/10.1371/journal.pone.0187967

Amberbir, A., Lin, S. H., Berman, J., Muula, A., Jacoby, D., Wroe, E., Maliwichi-Nyirenda, C., Mwapasa, V., Crampin, A., Makwero, M., Singogo, E., Phiri, S., Gordon, S., Tobe, S. W., Masiye, J., Newsome, B., Hosseinipour, M., Nyirenda, M. J., & van Oosterhout, J. J. (2019). Systematic Review of Hypertension and Diabetes Burden, Risk Factors, and Interventions for Prevention and Control in Malawi: The NCD BRITE Consortium. *Global Heart*, *14*(2), 109–118. https://doi.org/10.1016/j.gheart.2019.05.001

Apicella, M., Campopiano, M. C., Mantuano, M., Mazoni, L., Coppelli, A., & Del Prato, S. (2020). COVID-19 in people with diabetes: Understanding the reasons for worse outcomes. *The Lancet Diabetes & Endocrinology*, *8*(9), 782–792. https://doi.org/10.1016/S2213-8587(20)30238-2

Baichoo, S., Souilmi, Y., Panji, S., Botha, G., Meintjes, A., Hazelhurst, S., Bendou, H., Beste, E. de, Mpangase, P. T., Souiai, O., Alghali, M., Yi, L., O’Connor, B. D., Crusoe, M., Armstrong, D., Aron, S., Joubert, F., Ahmed, A. E., Mbiyavanga, M., …

Mulder, N. (2018). Developing reproducible bioinformatics analysis workflows for heterogeneous computing environments to support African genomics. *BMC Bioinformatics*, *19*, 457. https://doi.org/10.1186/s12859-018-2446-1

Bam, N. E., Mabunda, S. A., Ntsaba, J., Apalata, T., Nomatshila, S. C., & Chitha, W. (2020). The association between HIV tri-therapy with the development of Type-2 Diabetes Mellitus in a rural South African District: A case-control study. *PLOS ONE*, *15*(12), e0244067. https://doi.org/10.1371/journal.pone.0244067

Berkowitz, N., Okorie, A., Goliath, R., Levitt, N., Wilkinson, R. J., & Oni, T. (2018). The prevalence and determinants of active tuberculosis among diabetes patients in Cape Town, South Africa, a high HIV/TB burden setting. *Diabetes Research and Clinical Practice*, *138*, 16–25. https://doi.org/10.1016/j.diabres.2018.01.018

Bertram, M. Y., Jaswal, A. V. S., Van Wyk, V. P., Levitt, N. S., & Hofman, K. J. (2013). The non-fatal disease burden caused by type 2 diabetes in South Africa, 2009. *Global Health Action*, *6*(1), 19244. https://doi.org/10.3402/gha.v6i0.19244

Bigna, J. J., & Noubiap, J. J. (2019). The rising burden of non-communicable diseases in sub-Saharan Africa. *The Lancet Global Health*, *7*(10), e1295–e1296. https://doi.org/10.1016/S2214-109X(19)30370-5

Birungi, J., Kivuyo, S., Garrib, A., Mugenyi, L., Mutungi, G., Namakoola, I., Mghamba, J., Ramaiya, K., Wang, D., Maongezi, S., Musinguzi, J., Mugisha, K., Etukoit, B. M., Kakande, A., Niessen, L. W., Okebe, J., Shiri, T., Meshack, S., Lutale, J., … Jaffar, S. (2021). Integrating health services for HIV infection, diabetes and hypertension in sub-Saharan Africa: A cohort study. *BMJ Open*, *11*(11), e053412. https://doi.org/10.1136/bmjopen-2021-053412

Bosire, E. N. (2021). Patients' Experiences of Comorbid HIV/AIDS and Diabetes Care and Management in Soweto, South Africa. *Qualitative Health Research*, *31*(2), 373–384. https://doi.org/10.1177/1049732320967917

Boulle, A., Heekes, A., Tiffin, N., Smith, M., Mutemaringa, T., Zinyakatira, N., Phelanyane, F., Pienaar, C., Buddiga, K., Coetzee, E., Rooyen, R. van, Dyers, R., Fredericks, N., Loff, A., Shand, L., Moodley, M., Vega, I. de, & Vallabhjee, K. (2019). Data Centre Profile: The Provincial Health Data Centre of the Western Cape Province, South Africa. *International Journal of Population Data Science*, *4*(2), Article 2. https://doi.org/10.23889/ijpds.v4i2.1143

Bramante, C. T., Ingraham, N. E., Murray, T. A., Marmor, S., Hovertsen, S., Gronski, J., McNeil, C., Feng, R., Guzman, G., Abdelwahab, N., King, S., Tamariz, L., Meehan, T., Pendleton, K. M., Benson, B., Vojta, D., & Tignanelli, C. J. (2021). Metformin and risk of mortality in patients hospitalised with COVID-19: A retrospective cohort analysis. *The Lancet Healthy Longevity*, *2*(1), e34–e41. https://doi.org/10.1016/S2666-7568(20)30033-7

Brumpton, B. M., Graham, S., Surakka, I., Skogholt, A. H., Løset, M., Fritsche, L. G., Wolford, B., Zhou, W., Nielsen, J. B., Holmen, O. L., Gabrielsen, M. E., Thomas, L., Bhatta, L., Rasheed, H., Zhang, H., Kang, H. M., Hornsby, W., Moksnes, M. R., Coward, E., … Willer, C. J. (2022). The HUNT study: A population-based cohort for genetic research. *Cell Genomics*, *2*(10), 100193. https://doi.org/10.1016/j.xgen.2022.100193

Chen, G., Adeyemo, A., Zhou, J., Chen, Y., Huang, H., Doumatey, A., Lashley, K., Agyenim-Boateng, K., Eghan, B. A., Acheampong, J., Fasanmade, O., Johnson, T., Okafor, G., Oli, J., Amoah, A., & Rotimi, C. (2007). Genome-wide search for susceptibility genes to type 2 diabetes in West Africans: Potential role of C-peptide. *Diabetes Research and Clinical Practice*, *78*(3), e1–e6. https://doi.org/10.1016/j.diabres.2007.04.010

Chen, J., Spracklen, C. N., Marenne, G., Varshney, A., Corbin, L. J., Luan, J., Willems, S. M., Wu, Y., Zhang, X., Horikoshi, M., Boutin, T. S., Mägi, R., Waage, J., Pitsilides, A., Li-Gao, R., Chan, K. H. K., Yao, J., Anasanti, M. D., Chu, A. Y., … Barroso, I. (2021). The Trans-Ancestral Genomic Architecture of Glycemic Traits. *Nature Genetics*, *53*(6), 840–860. https://doi.org/10.1038/s41588-021-00852-9

Chen, J., Sun, M., Adeyemo, A., Pirie, F., Carstensen, T., Pomilla, C., Doumatey, A. P., Chen, G., Young, E. H., Sandhu, M., Morris, A. P., Barroso, I., McCarthy, M. I., Mahajan, A., Wheeler, E., Rotimi, C. N., & Motala, A. A. (2019). Genome-wide association study of type 2 diabetes in Africa. *Diabetologia*, *62*(7), 1204–1211. https://doi.org/10.1007/s00125-019-4880-7

Chikowore, T., Ekoru, K., Vujkovi, M., Gill, D., Pirie, F., Young, E., Sandhu, M. S., McCarthy, M., Rotimi, C., Adeyemo, A., Motala, A., & Fatumo, S. (2022). Polygenic Prediction of Type 2 Diabetes in Africa. *Diabetes Care*, *45*(3), 717–723. https://doi.org/10.2337/dc21-0365

Chimusa, E. R., Daya, M., Möller, M., Ramesar, R., Henn, B. M., Helden, P. D. van, Mulder, N. J., & Hoal, E. G. (2013). Determining Ancestry Proportions in Complex Admixture Scenarios in South Africa Using a Novel Proxy Ancestry Selection Method. *PLOS ONE*, *8*(9), e73971. https://doi.org/10.1371/journal.pone.0073971

Choudhury, A., Aron, S., Sengupta, D., Hazelhurst, S., & Ramsay, M. (2018). African genetic diversity provides novel insights into evolutionary history and local adaptations. *Human Molecular Genetics*, *27*(R2), R209–R218. https://doi.org/10.1093/hmg/ddy161

Choudhury, A., Ramsay, M., Hazelhurst, S., Aron, S., Bardien, S., Botha, G., Chimusa, E. R., Christoffels, A., Gamieldien, J., Sefid-Dashti, M. J., Joubert, F., Meintjes, A., Mulder, N., Ramesar, R., Rees, J., Scholtz, K., Sengupta, D., Soodyall, H., Venter, P., … Pepper, M. S. (2017). Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. *Nature Communications*, *8*(1), Article 1. https://doi.org/10.1038/s41467-017-00663-9

Choudhury, A., Sengupta, D., Ramsay, M., & Schlebusch, C. (2021). Bantu-speaker migration and admixture in southern Africa. *Human Molecular Genetics*, *30*(R1), R56–R63. https://doi.org/10.1093/hmg/ddaa274

Christoffels, A., & Abayomi, A. (2020). Careful governance of African biobanks. *The Lancet*, *395*(10217), 29–30. https://doi.org/10.1016/S0140-6736(19)32624-8

Croxton, T., Swanepoel, C., Musinguzi, H., Kader, M., Ozumba, P., Pillay, A.-D., Syed, R., Swartz, G., Kyobe, S., Lwanga, N., Katabazi, F. A., Edgar, K., Ndidi, A., Jonathan, E., Onyemata, E., Isaacs, S., Mayne, E. S., Joloba, M., Jentsch, U., … Penno, S. (2017). Lessons Learned from Biospecimen Shipping Among the Human Heredity and Health in Africa Biorepositories. *Biopreservation and Biobanking*, *15*(2), 103–110. https://doi.org/10.1089/bio.2017.0009

Daya, M., Merwe, L. van der, Galal, U., Möller, M., Salie, M., Chimusa, E. R., Galanter, J. M., Helden, P. D. van, Henn, B. M., Gignoux, C. R., & Hoal, E. (2013). A Panel of Ancestry Informative Markers for the Complex Five-Way Admixed South African Coloured Population. *PLOS ONE*, *8*(12), e82224. https://doi.org/10.1371/journal.pone.0082224

de Vries, J., Munung, S. N., Matimba, A., McCurdy, S., Ouwe Missi Oukem-Boyer, O., Staunton, C., Yakubu, A., Tindana, P., & the H3Africa Consortium. (2017). Regulation of genomic and biobanking research in Africa: A content analysis of ethics guidelines, policies and procedures from 22 African countries. *BMC Medical Ethics*, *18*(1), 8. https://doi.org/10.1186/s12910-016-0165-6

de Wit, E., Delport, W., Rugamika, C. E., Meintjes, A., Möller, M., van Helden, P. D., Seoighe, C., & Hoal, E. G. (2010). Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape. *Human Genetics*, *128*(2), 145–153. https://doi.org/10.1007/s00439-010-0836-1

DeForest, N., & Majithia, A. R. (2022). Genetics of Type 2 Diabetes: Implications from Large-Scale Studies. *Current Diabetes Reports*, *22*(5), 227–235. https://doi.org/10.1007/s11892-022-01462-3

Fatumo, S. (2020). The opportunity in African genome resource for precision medicine. *EBioMedicine*, *54*, 102721. https://doi.org/10.1016/j.ebiom.2020.102721

Garrib, A., Birungi, J., Lesikari, S., Namakoola, I., Njim, T., Cuevas, L., Niessen, L., Mugisha, K., Mutungi, G., Mghamba, J., Ramaiya, K., Jaffar, S., Mfinanga, S., & Nyirenda, M. (2019). Integrated care for human immunodeficiency virus, diabetes and hypertension in Africa. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, *113*(12), 809–812. https://doi.org/10.1093/trstmh/try098

Gennaro, F. D., Marotta, C., Antunes, M., & Pizzol, D. (2019). Diabetes in active tuberculosis in low-income countries: To test or to take care? *The Lancet Global Health*, *7*(6), e707. https://doi.org/10.1016/S2214-109X(19)30173-1

Gill, G. V., Mbanya, J.-C., Ramaiya, K. L., & Tesfaye, S. (2008). A sub-Saharan African perspective of diabetes. *Diabetologia*, *52*(1), 8. https://doi.org/10.1007/s00125-008-1167-9

Gouda, H. N., Charlson, F., Sorsdahl, K., Ahmadzada, S., Ferrari, A. J., Erskine, H., Leung, J., Santamauro, D., Lund, C., Aminde, L. N., Mayosi, B. M., Kengne, A. P., Harris, M., Achoki, T., Wiysonge, C. S., Stein, D. J., & Whiteford, H. (2019). Burden of non-communicable diseases in sub-Saharan Africa, 1990–2017: Results from the Global Burden of Disease Study 2017. *The Lancet Global Health*, *7*(10), e1375–e1387. https://doi.org/10.1016/S2214-109X(19)30374-2

Grady, C., Eckstein, L., Berkman, B., Brock, D., Cook-Deegan, R., Fullerton, S. M., Greely, H., Hansson, M. G., Hull, S., Kim, S., Lo, B., Pentz, R., Rodriguez, L., Weil, C., Wilfond, B. S., & Wendler, D. (2015). Broad Consent For Research With Biological Samples: Workshop Conclusions. *The American Journal of Bioethics : AJOB*, *15*(9), 34–42. https://doi.org/10.1080/15265161.2015.1062162

Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M. O., Choudhury, A., Ritchie, G. R. S., Xue, Y., Asimit, J., Nsubuga, R. N., Young, E. H., Pomilla, C., Kivinen, K., Rockett, K., Kamali, A., … Sandhu, M. S. (2015). The African Genome Variation Project shapes medical genetics in Africa. *Nature*, *517*(7534), 327–332. https://doi.org/10.1038/nature13997

Hall, V., Thomsen, R. W., Henriksen, O., & Lohse, N. (2011). Diabetes in Sub Saharan Africa 1999-2011: Epidemiology and public health implications. a systematic review. *BMC Public Health*, *11*(1), 564. https://doi.org/10.1186/1471-2458-11-564

IDF Diabetes Atlas 10th Edition. (n.d.). Retrieved May 2, 2022, from https://diabetesatlas.org/data/

Irgam, K., Reddy, B. S., Hari, S. G., Banapuram, S., & Reddy, B. M. (2021). The genetic susceptibility profile of type 2 diabetes and reflection of its possible role related to reproductive dysfunctions in the southern Indian population of Hyderabad. *BMC Medical Genomics*, *14*(1), 272. https://doi.org/10.1186/s12920-021-01129-0

Ishigaki, K., Akiyama, M., Kanai, M., Takahashi, A., Kawakami, E., Sugishita, H., Sakaue, S., Matoba, N., Low, S.-K., Okada, Y., Terao, C., Amariuta, T., Gazal, S., Kochi, Y., Horikoshi, M., Suzuki, K., Ito, K., Koyama, S., Ozaki, K., … Kamatani, Y. (2020). Large scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases. *Nature Genetics*, *52*(7), 669–679. https://doi.org/10.1038/s41588-020-0640-3

Jaffar, S., Ramaiya, K., Karekezi, C., Sewankambo, N., & RESPOND-AFRICA Group. (2021). Controlling diabetes and hypertension in sub-Saharan Africa: Lessons from HIV programmes. *Lancet (London, England)*, *398*(10306), 1111–1113. https://doi.org/10.1016/S0140-6736(21)01731-1

Jongeneel, C. V., Kotze, M. J., Bhaw-Luximon, A., Fadlelmola, F. M., Fakim, Y. J., Hamdi, Y., Kassim, S. K., Kumuthini, J., Nembaware, V., Radouani, F., Tiffin, N., & Mulder, N. (2022). A View on Genomic Medicine Activities in Africa: Implications for Policy. *Frontiers in Genetics*, *13*. https://www.frontiersin.org/articles/10.3389/fgene.2022.769919

Kamiza, A. B., Toure, S. M., Vujkovic, M., Machipisa, T., Soremekun, O. S., Kintu, C., Corpas, M., Pirie, F., Young, E., Gill, D., Sandhu, M. S., Kaleebu, P., Nyirenda, M., Motala, A. A., Chikowore, T., & Fatumo, S. (2022). Transferability of genetic risk scores in African populations. *Nature Medicine*, *28*(6), 1163–1166. https://doi.org/10.1038/s41591-022-01835-x

Kulminski, A. M., Loika, Y., Culminskaya, I., Arbeev, K. G., Ukraintseva, S. V., Stallard, E., & Yashin, A. I. (2016). Explicating heterogeneity of complex traits has strong potential for improving GWAS efficiency. *Scientific Reports*, *6*(1), Article 1. https://doi.org/10.1038/srep35390

Levitt, N. S. (2008). Diabetes in Africa: Epidemiology, management and healthcare challenges. *Heart*, *94*(11), 1376–1382. https://doi.org/10.1136/hrt.2008.147306

Levitt, N. S., Peer, N., Steyn, K., Lombard, C., Maartens, G., Lambert, E. V., & Dave, J. A. (2016). Increased risk of dysglycaemia in South Africans with HIV; especially those on protease inhibitors. *Diabetes Research and Clinical Practice*, *119*, 41–47. https://doi.org/10.1016/j.diabres.2016.03.012

Maloy, J. W., & Bass, P. F. (2020). Understanding Broad Consent. *The Ochsner Journal*, *20*(1), 81–86. https://doi.org/10.31486/toj.19.0088

Manyema, M., Veerman, J. L., Chola, L., Tugendhaft, A., Labadarios, D., & Hofman, K. (2015). Decreasing the Burden of Type 2 Diabetes in South Africa: The Impact of Taxing Sugar-Sweetened Beverages. *PLOS ONE*, *10*(11), e0143050. https://doi.org/10.1371/journal.pone.0143050

McGurnaghan, S. J., Weir, A., Bishop, J., Kennedy, S., Blackbourn, L. A. K., McAllister, D. A., Hutchinson, S., Caparrotta, T. M., Mellor, J., Jeyam, A., O’Reilly, J. E., Wild, S. H., Hatam, S., Höhn, A., Colombo, M., Robertson, C., Lone, N., Murray, J., Butterly, E., … McCoubrey, J. (2021). Risks of and risk factors for COVID-19 disease in people with diabetes: A cohort study of the total population of Scotland. *The Lancet Diabetes & Endocrinology*, *9*(2), 82–93. https://doi.org/10.1016/S2213-8587(20)30405-8

Mikkelsen, R. B., Gjerris, M., Waldemar, G., & Sandøe, P. (2019). Broad consent for biobanks is best – provided it is also deep. *BMC Medical Ethics*, *20*(1), 71. https://doi.org/10.1186/s12910-019-0414-6

Mulder, N. (2017). Development to enable precision medicine in Africa. *Personalized Medicine*, *14*(6), 467–470. https://doi.org/10.2217/pme-2017-0055

Mulder, N., Abimiku, A., Adebamowo, S. N., de Vries, J., Matimba, A., Olowoyo, P., Ramsay, M., Skelton, M., & Stein, D. J. (2018). H3Africa: Current perspectives.

Pharmacogenomics and Personalized Medicine, *11*, 59–66. https://doi.org/10.2147/PGPM.S141546

Mulder, N. J., Adebiyi, E., Adebiyi, M., Adeyemi, S., Ahmed, A., Ahmed, R., Akanle, B., Alibi, M., Armstrong, D. L., Aron, S., Ashano, E., Baichoo, S., Benkahla, A., Brown, D. K., Chimusa, E. R., Fadlelmola, F. M., Falola, D., Fatumo, S., Ghedira, K., … H3ABioNet Consortium, as members of the H3Africa Consortium. (2017). Development of Bioinformatics Infrastructure for Genomics Research. *Global Heart*, *12*(2), 91–98. https://doi.org/10.1016/j.gheart.2017.01.005

Nagai, A., Hirata, M., Kamatani, Y., Muto, K., Matsuda, K., Kiyohara, Y., Ninomiya, T., Tamakoshi, A., Yamagata, Z., Mushiroda, T., Murakami, Y., Yuji, K., Furukawa, Y., Zembutsu, H., Tanaka, T., Ohnishi, Y., Nakamura, Y., & Kubo, M. (2017). Overview of the BioBank Japan Project: Study design and profile. *Journal of Epidemiology*, *27*(3 Suppl), S2–S8. https://doi.org/10.1016/j.je.2016.12.005

Nembaware, V., Johnston, K., Diallo, A. A., Kotze, M. J., Matimba, A., Moodley, K., Tangwa, G. B., Torrorey-Sawe, R., & Tiffin, N. (2019). A framework for tiered informed consent for health genomic research in Africa. *Nature Genetics*, *51*(11), 1566–1571. https://doi.org/10.1038/s41588-019-0520-x

Pastakia, S. D., Pekny, C. R., Manyara, S. M., & Fischer, L. (2017). Diabetes in sub-Saharan Africa – from policy to practice to progress: Targeting the existing gaps for future care for diabetes. *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*, *10*, 247–263. https://doi.org/10.2147/DMSO.S126314

Patin, E., Lopez, M., Grollemund, R., Verdu, P., Harmant, C., Quach, H., Laval, G., Perry, G. H., Barreiro, L. B., Froment, A., Heyer, E., Massougbodji, A., Fortes-Lima, C., Migot-Nabias, F., Bellis, G., Dugoujon, J.-M., Pereira, J. B., Fernandes, V., Pereira, L., … Quintana-Murci, L. (2017). Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science*, *356*(6337), 543–546. https://doi.org/10.1126/science.aal1988

Petersen, D. C., Libiger, O., Tindall, E. A., Hardie, R.-A., Hannick, L. I., Glashoff, R. H., Mukerji, M., Consortium, I. G. V., Fernandez, P., Haacke, W., Schork, N. J., & Hayes, V. M. (2013). Complex Patterns of Genomic Admixture within Southern Africa. *PLOS Genetics*, *9*(3), e1003309. https://doi.org/10.1371/journal.pgen.1003309

Policy Paper: A Framework for the Implementation of Genomic Medicine for Public Health in Africa | The AAS. (n.d.). Retrieved January 25, 2023, from https://www.aasciences.africa/publications/policy-paper-framework-implementation-genomic-medicine-public-health-africa

Ramsay, M. (2012). Africa: Continent of genome contrasts with implications for biomedical research and health. *FEBS Letters*, *586*(18), 2813–2819. https://doi.org/10.1016/j.febslet.2012.07.061

Rotimi, C. N., Chen, G., Adeyemo, A. A., Furbert-Harris, P., Guass, D., Zhou, J., Berg, K., Adegoke, O., Amoah, A., Owusu, S., Acheampong, J., Agyenim-Boateng, K., Eghan, B. A., Jr., Oli, J., Okafor, G., Ofoegbu, E., Osotimehin, B., Abbiyesuku, F., Johnson, T., … Collins, F. S. (2004). A Genome-Wide Search for Type 2 Diabetes Susceptibility Genes in West Africans: The Africa America Diabetes Mellitus (AADM) Study. *Diabetes*, *53*(3), 838–841. https://doi.org/10.2337/diabetes.53.3.838

Spracklen, C. N., Chen, P., Kim, Y. J., Wang, X., Cai, H., Li, S., Long, J., Wu, Y., Wang, Y. X., Takeuchi, F., Wu, J.-Y., Jung, K.-J., Hu, C., Akiyama, K., Zhang, Y., Moon, S., Johnson, T. A., Li, H., Dorajoo, R., … Sim, X. (2017). Association analyses of East Asian individuals and trans-ancestry analyses with European individuals reveal new loci associated with cholesterol and triglyceride levels. *Human Molecular Genetics*, *26*(9), 1770–1784. https://doi.org/10.1093/hmg/ddx062

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine*, *12*(3), e1001779. https://doi.org/10.1371/journal.pmed.1001779

Teo, Y.-Y., Small, K. S., & Kwiatkowski, D. P. (2010). Methodological challenges of genome-wide association analysis in Africa. *Nature Reviews Genetics*, *11*(2), Article 2. https://doi.org/10.1038/nrg2731

Tiffin, N. (2018). Tiered informed consent: Respecting autonomy, agency and individuality in Africa. *BMJ Global Health*, *3*(6), e001249. https://doi.org/10.1136/bmjgh-2018-001249

Todd, J., & Mahande, M. J. (2020). Editorial: The Use of Routine Health Data in Low- and Middle-Income Countries. *Frontiers in Public Health*, *8*. https://www.frontiersin.org/article/10.3389/fpubh.2020.00413

Wabiri, N., Naidoo, I., Mungai, E., Samuel, C., & Ngwenya, T. (2019). The Arts and Tools for Using Routine Health Data to Establish HIV High Burden Areas: The Pilot Case of KwaZulu-Natal South Africa. *Frontiers in Public Health*, *7*. https://www.frontiersin.org/articles/10.3389/fpubh.2019.00335

Wonkam, A. (2021). Sequence three million genomes across Africa. *Nature*, *590*(7845), 209–211. https://doi.org/10.1038/d41586-021-00313-7

Yako, Y. Y., Guewo-Fokeng, M., Balti, E. V., Bouatia-Naji, N., Matsha, T. E., Sobngwi, E., Erasmus, R. T., Echouffo-Tcheugui, J. B., & Kengne, A. P. (2016). Genetic risk of type 2 diabetes in populations of the African continent: A systematic review and meta-analyses. *Diabetes Research and Clinical Practice*, *114*, 136–150. https://doi.org/10.1016/j.diabres.2016.01.003

2. Chapter 2 : An e-Consent framework for tiered informed consent for human genomic research in the global South, implemented as a REDCap template

Tamuhla T, Tiffin N, Allie T. An e-consent framework for tiered informed consent for human genomic research in the global south, implemented as a REDCap template. BMC Med Ethics. 2022;23: 119. doi:10.1186/s12910-022-00860-2

Author contributions

TT, NT, and TA designed the project and generated the content. TT and TA implemented the REDCap template. TT wrote the first draft of the manuscript. NT, TT and TA finalised the manuscript.

Relevance of paper to the thesis

This chapter addressed objective 1 of optimising and adopting a tiered informed consent that is suitable for both the cohort study design and for conducting human genomic research in the African context. Since creating a virtual genotyped cohort involves asking for sensitive genomic data and access to health records, special attention should be given to interactions with the participants and the consent process, to ensure that any consent given is truly informed. Traditional consent has largely been broad consent, but with introduction of legislation to protect personal information such as the POPI act in South Africa, there have been calls to move to a tiered consent model in research. At the same time there has been a shift to more collaborative research and data sharing among researchers. Therefore, to foster ethical data sharing and collaboration, there is a need to simplify and standardise the process through which researchers can identify consenting participants. At the same time there is also a need to ensure that participants give consent that is truly informed especially in genomic research where there is a privacy risk even with deidentified data. To address both these needs, we designed an electronic consent framework for tiered informed consent based in REDCap.

2.1. Abstract

Research involving human participants requires their consent, and it is common practice to capture consent information on paper and store those hard copies, presenting issues such as long-term storage requirements, inefficient retrieval of consent forms for reference or future use, and the potential for transcription errors when transcribing captured informed consent. There have been calls to move to electronic capture of the consent provided by research participants (e-consent) as a way of addressing these issues.

A tiered framework for e-consent was designed using the freely available features in the inbuilt REDCap e-consent module. We implemented ‘branching logic’, ‘wet signature’ and ‘auto-archiver’ features to the main informed consent and withdrawal of consent documents. The branching logic feature streamlines the consent process by making follow-up information available depending on participant response, the ‘wet signature’ feature enables a timestamped electronic signature to be appended to the e-consent documents and the ‘auto-archiver’ allows for PDF copies of the e-consent documents to be stored in the database. When designing the content layout, we provided example participant information text which can be modified as required. Emphasis was placed on the flow of information to optimise participant understanding and this was achieved by merging the consent and participant information into one document where the consent questions were asked immediately after the corresponding participant information. In addition, we have provided example text for a generic human genomic research study, which can be easily edited and modified according to specific requirements.

Building informed consent protocols and forms without prior experience can be daunting, so we have provided researchers with a REDCap template that can be directly incorporated into REDCap databases. It prompts researchers about the types of consent they can request for genomics studies and assists them with suggestions for the language they might use for participant information and consent questions. The use of this tiered e-consent module can ensure the accurate and efficient electronic capture and storage of the consents given by participants in a format that can be easily queried and can thus facilitate ethical and effective onward sharing of data and samples whilst upholding individual participant preferences.

2.2. Background

Research involving human subjects generally requires voluntary participation and signed consent from participants granting researchers permission to use their biological and/or health data (Artal & Rubenfeld, 2017; Dankar et al., 2019; Pranati, 2010). To facilitate this process, researchers are required to provide detailed and transparent information about their research in a format that allows eligible participants to make informed decisions about whether to volunteer to participate in the research (Nishimura et al., 2013; Pranati, 2010; Tiffin, 2018; Trinidad et al., 2012). The informed consent process has frequently been criticised for not being participant-centred but rather more focused on meeting legal and regulatory requirements resulting in consent forms which use complex technical terms which lay persons cannot understand - especially those in vulnerable populations with limited health literacy (Nishimura et al., 2013; Pranati, 2010; Tiffin, 2018). There is therefore a need to improve the informed consent process by using consent documents that are straightforward and use easy-to-understand language to ensure that participants give truly informed consent (Lentz et al., 2016; Nishimura et al., 2013; Pranati, 2010).

It is common practice to capture consent information on paper and store those hard copies, and while this has its advantages, it presents issues such as long-term storage requirements and inefficient retrieval of consent forms for reference or future use (De Sutter et al., 2021; Lentz et al., 2016; Vanaken & Masand, 2019). In addition, for tiered informed consent where participants answer a variety of questions about allowed data or specimen us, paper-based consents are inefficient and impractical for determining whose data or which specific data elements have consent for onward sharing, meta-analyses or sharing in aggregated form (Chalil Madathil et al., 2013). Participants may also express a variety of preferences for future contact and/or feedback of findings from the research programme. While it is possible to transcribe this information from hard copies into electronic format, this is time-consuming and prone to data capture error, which might lead to unacceptable transgression of participants' choices about how their data and specimens might be used (Tiffin, 2018). There have been calls to move to electronic capture of the consent process (e-consent) as a way of addressing these issues (Chalil Madathil et al., 2013; De Sutter et al., 2020; Lentz et al., 2016; Vanaken & Masand, 2019). However, there has been slow uptake of e-consent because of technical, legislative, and regulatory barriers to setting up and implementing e-consent platforms. These include concerns about data security, legal validity of electronic signatures, and initial development costs (De Sutter et al., 2020, 2021; Haussen et al., 2017; Vanaken & Masand, 2019).

Researchers can find designing informed consent processes overwhelming and may not know how to implement them or what content is required. Using our experience in conducting tiered informed consent in South Africa, we have designed a REDCap-based electronic tiered informed consent framework that can aid in reducing barriers to uptake and implementation of e-consent in low- and middle-income countries. The framework is designed to improve the informed consent process for both participants and researchers involved in human genomic research, firstly by providing a comprehensive list of information for researchers to include in the consent documents, thus providing a tool which they can use as a check list to ensure that all essential information is available; and secondly providing researchers with ready-to-use, downloadable template consent documents which have been written in straightforward genomic research language that is easier for participants to understand. In this paper we present the content for the modules that can be used to construct the integrated participant information and consent form and describe how the REDCap template can be implemented to create study-specific tiered consent. A checklist that summarises the processes and consent modules is provided as Supplementary file 2.1.

The REDCap-based tiered e-consent module presented here facilitates the electronic capture of participants' consent choices so no additional data entry is required and errors are kept to a minimum. We recommend that this process is undertaken by trained personnel who can accurately capture the preferences of the participants. We have also provided a REDCap database template so that researchers can easily incorporate this tiered informed consent module into their new REDCap research databases in a "ready-to-use" format, selecting elements and modifying the contents to fulfil their requirements without needing to develop new material *de novo*. For re-use of data and specimens, the captured information can be rapidly and easily queried to identify which resources have consent for other onward uses, and which participants might be re-contacted in the future for follow-up or related studies – thus facilitating efficient and ethical data-sharing and follow-up with participants where their consent has been given.

2.3. Construction and content

2.3.1. Setting up the tiered e-consent framework in REDCap

Research Electronic Data Capture (REDCap) is a secure online databasing platform that allows production of generalisable data capture instruments for research (Harris et al., 2009). REDCap has an inbuilt e-consent framework where consent is administered as a survey (Lawrence et al., 2020). The tiered e-consent framework for genomic research was designed using tools in REDCap version 10.9.4 and is available as template data dictionary (ConsentFramework_Data_Dictionary) which researchers can download from GitHub (https://github.com/CIDRI-Africa/e-Consent-framework) and import into REDCap to set-up their own tiered e-consent module (Figure 2.1).

The tiered e-consent template is modular, allowing users to select elements which are suitable for their study. In addition, guidance documents which include a REDCap set-up guide, an instrument index which describes the data collection instruments available in the module and PDF copies of example data collection instruments are also available in the GitHub repository (see Supplementary file 2.2).

All consent documents used in a human genomics study need ethics approval before they can be used. The ethics review process for this REDCap based tiered e-consent module is like that of the paper-based consent because it does not contain any multi-media information such as videos. All the online tiered e-consent documents can be downloaded and submitted as PDFs to the institutional review board (IRB) and if required, a link can be set-up to give the IRB access to consent surveys online, on the REDCap platform. To successfully implement the tiered e-consent, research staff will need to be trained on how to navigate the REDCap platform and how to administer tiered informed e-consent.

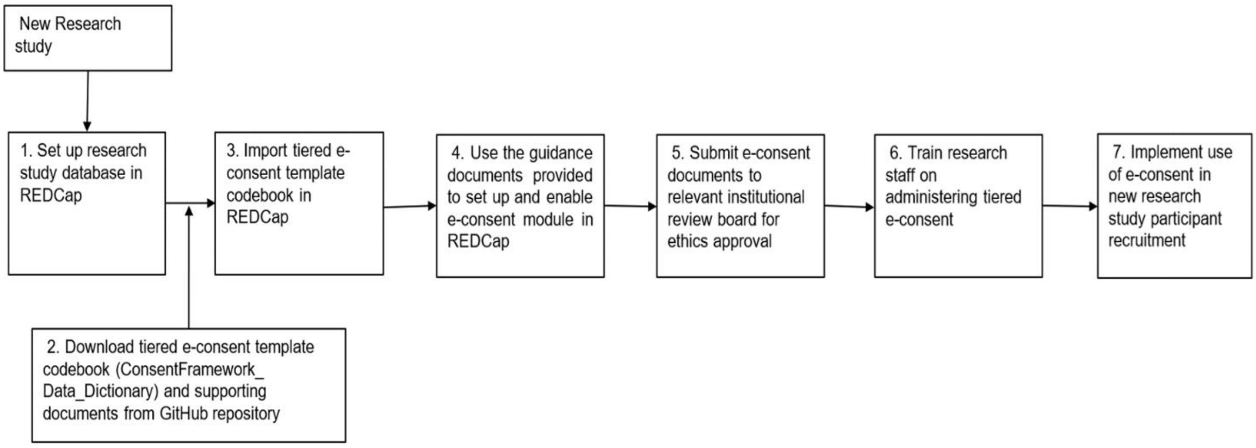

New Research study
1. Set up research study database in REDCap
2. Download tiered e-consent template codebook (ConsentFramework_ Data_Dictionary) and supporting documents from GitHub repository
3. Import tiered e-consent template codebook in REDCap
4. Use the guidance documents provided to set up and enable e-consent module in REDCap
5. Submit e-consent documents to relevant institutional review board for ethics approval
6. Train research staff on administering tiered e-consent
7. Implement use of e-consent in new research study participant recruitment

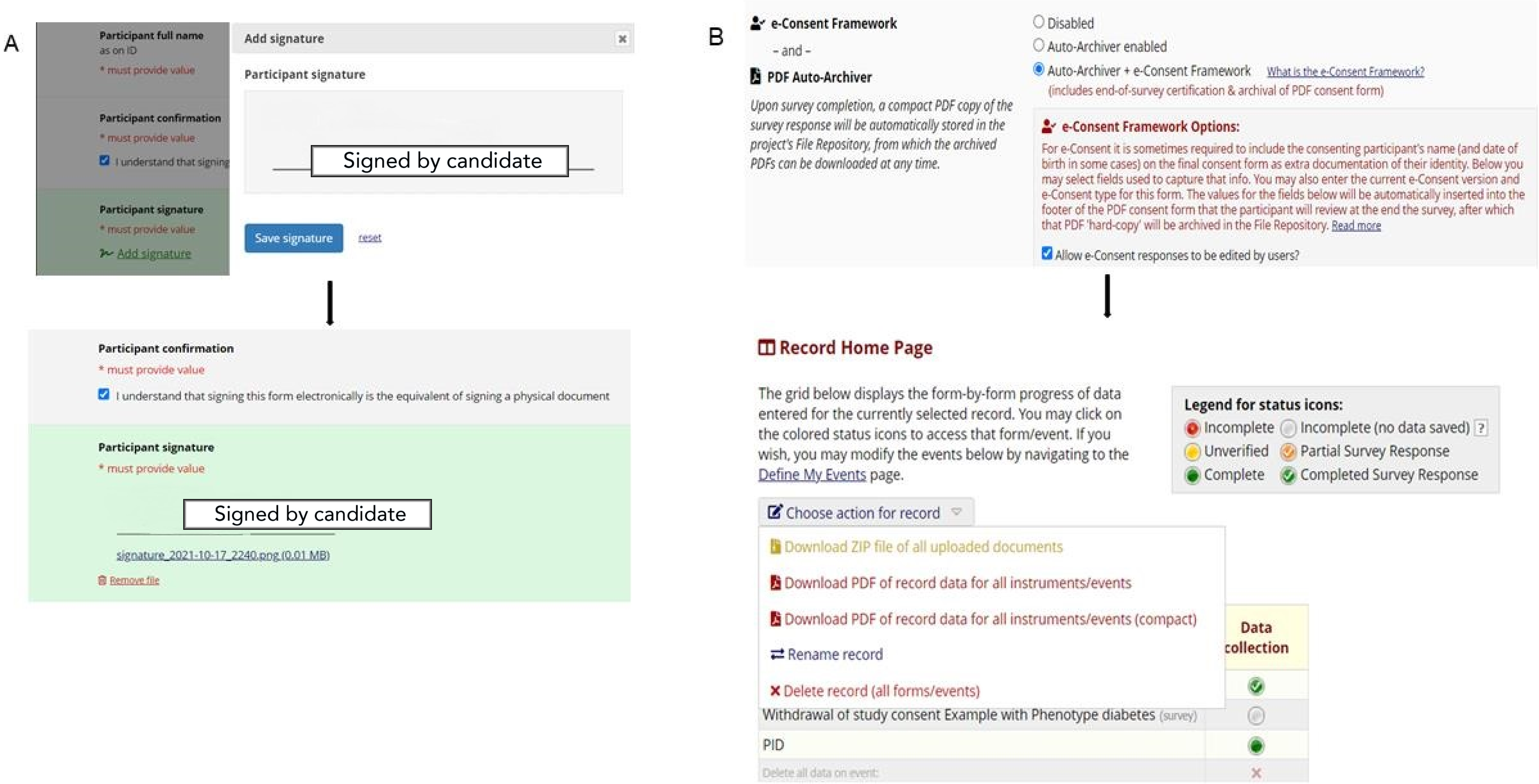
Figure 2.1. Flow diagram showing the workflow for setting up and implementing the tiered e-consent framework in REDCap for a new human genomic research study.

2.3.2. Tiered e-consent data collection instruments

The tiered e-consent module has three data capture instruments documents namely, the main consent and withdrawal of consent which are both surveys and an optional study meta data form. The inbuilt REDCap e-consent module has eight freely available features previously described by Lawrence *et al.*(2020) (Lawrence et al., 2020) which enhance the utility and security of the data capture instruments. For this tiered e-consent module we implemented ‘branching logic’, ‘wet signature’ and ‘PDF-consent document repository (auto-archiver)’ (Lawrence et al., 2020) to the main consent and withdrawal of consent documents. The branching logic feature streamlines the consent process by making follow-up information available depending on participant response and it was used in both the consent surveys. The ‘wet signature’ feature enables a timestamped electronic signature to be appended to the e-consent documents and the ‘auto-archiver’ allows for PDF copies of the e-consent documents to be stored in the database (Figure 2.2) (Lawrence et al., 2020).

2.3.3. Additional REDCap survey customisations

To ensure that tiered e-consent framework facilitates improved data quality, storage, retrieval and integrity, REDCap has additional customisations (see Supplementary file 2.3) which can be enabled for the consent surveys. When generating new records, the ‘*designate a secondary unique field*’ customisation allows the user to assign one of the variables such as the participant study ID as a unique value which cannot be duplicated. When this feature is enabled, each time that variable is entered, it is checked in real time to ensure that it has not been assigned already. This will help with improving data quality as participants will not be assigned the same study ID especially in multi-site studies or where multiple people are carrying out consent simultaneously. When enabled, the ‘*display the Today/Now button for all date and time fields on forms/surveys*’ ensures that the current date or time will be set automatically by clicking a button. The ‘*set a custom record label’* feature allows another variable such as the participant study ID to be appended to the system generated numeric record name, to simplify the query and retrieval of individual participant records from the database. To ensure data integrity, three additional customisations namely ‘*require a reason when making changes to existing records*’, ‘*enable the data history pop up for all data collection documents’* and ‘*enable the field comment log or data resolution workflow (data queries)*’ can be enabled. These features described in detail in Supplementary file 2.3, ensure that any changes made to the consent documents after verification and signing are not only sanctioned but are recorded appropriately to ensure data integrity. In addition, user rights and permissions can also be set to determine who can add and/or edit records in the tiered e-consent framework.



Participant full name
as on ID
* must provide value

Participant confirmation
* must provide value
☑ I understand that signing

Participant signature
* must provide value
Add signature

Add signature ✖

Participant signature

Signed by candidate

Save signature reset

Participant confirmation
* must provide value
☑ I understand that signing this form electronically is the equivalent of signing a physical document

Participant signature
* must provide value

Signed by candidate

signature_2021-10-17_2240.png (0.01 MB)

Remove file

e-Consent Framework
– and –
PDF Auto-Archiver

Upon survey completion, a compact PDF copy of the survey response will be automatically stored in the project's File Repository, from which the archived PDFs can be downloaded at any time.

○ Disabled
○ Auto-Archiver enabled
● Auto-Archiver + e-Consent Framework What is the e-Consent Framework?
(includes end-of-survey certification & archival of PDF consent form)

e-Consent Framework Options:

For e-Consent it is sometimes required to include the consenting participant's name (and date of birth in some cases) on the final consent form as extra documentation of their identity. Below you may select fields used to capture that info. You may also enter the current e-Consent version and e-Consent type for this form. The values for the fields below will be automatically inserted into the footer of the PDF consent form that the participant will review at the end the survey, after which that PDF 'hard-copy' will be archived in the File Repository. Read more

☑ Allow e-Consent responses to be edited by users?

Record Home Page

The grid below displays the form-by-form progress of data entered for the currently selected record. You may click on the colored status icons to access that form/event. If you wish, you may modify the events below by navigating to the Define My Events page.

Legend for status icons:
● Incomplete ○ Incomplete (no data saved) ?
● Unverified ✓ Partial Survey Response
● Complete ✓ Completed Survey Response

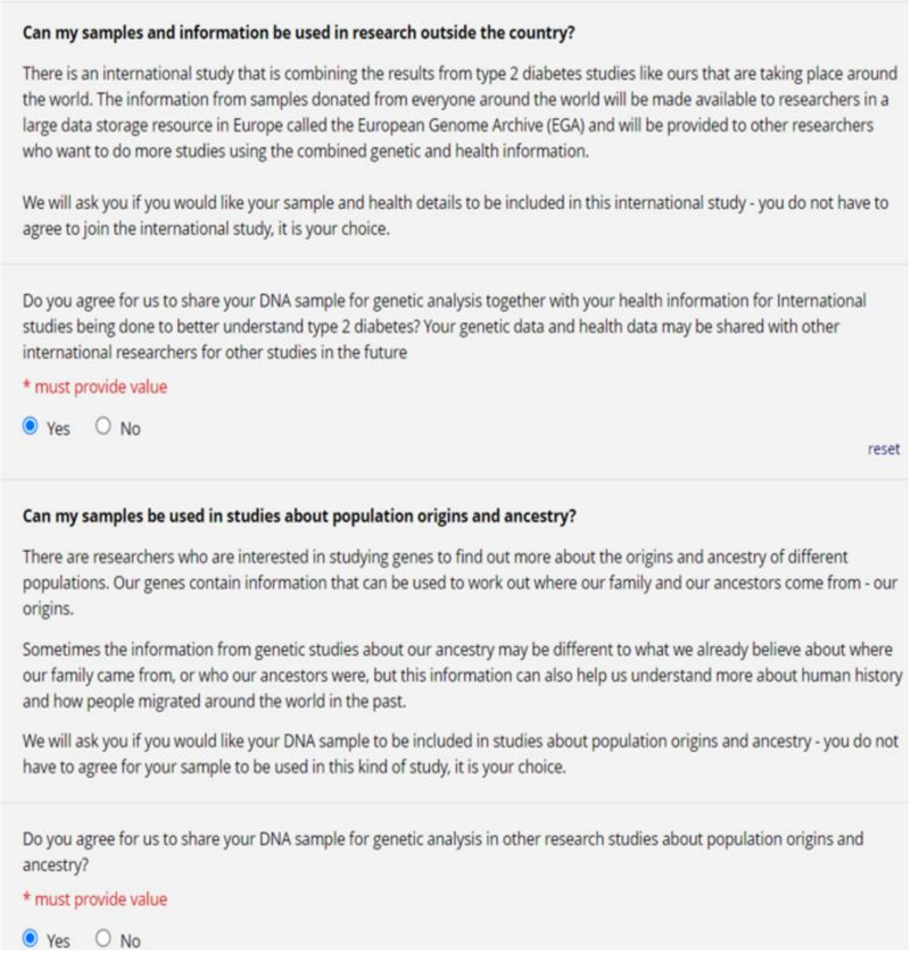
Choose action for record
- Download ZIP file of all uploaded documents
- Download PDF of record data for all instruments/events
- Download PDF of record data for all instruments/events (compact)
- Rename record
- ✕ Delete record (all forms/events)

	Data collection
	✓
Withdrawal of study consent Example with Phenotype diabetes (survey)	○
PID	●
Delete all data on event:	✕

Figure 2.2. The (A) 'wet signature' and (B) 'auto-archiver' features in the tiered e-consent framework

2.3.4. Language and layout of data collection instruments

The content of the main consent was adapted from the tiered informed consent framework of Nembware and colleagues (Nembaware et al., 2019) with some modifications, most notable of which was the addition of consent for the use of participant genomic data in population and ancestry studies which was excluded from that framework. Most consent documents have the same layout, where participant information is presented first, and consent questions are asked at the end or even on a separate document. When designing the content layout of the main consent document in the tiered e-consent framework, emphasis was placed on the flow of information to optimise participant understanding. This was achieved by merging the consent and participant information into one document where the consent questions were asked immediately after the corresponding participant information (Figure 2.3). This format is intended to allow the participant to ask further questions and seek clarity before making a consent choice. In addition, we have provided example text for a generic human genomic research study on type 2 diabetes where the researchers are collecting both DNA and routine electronic health data from participants (Figure 2.3). This text can be easily edited and modified to suit different research topics, and we recommend participant information and consent questions are modified and validated to suit each context in which this template is used.



Can my samples and information be used in research outside the country?

There is an international study that is combining the results from type 2 diabetes studies like ours that are taking place around the world. The information from samples donated from everyone around the world will be made available to researchers in a large data storage resource in Europe called the European Genome Archive (EGA) and will be provided to other researchers who want to do more studies using the combined genetic and health information.

We will ask you if you would like your sample and health details to be included in this international study - you do not have to agree to join the international study, it is your choice.

Do you agree for us to share your DNA sample for genetic analysis together with your health information for International studies being done to better understand type 2 diabetes? Your genetic data and health data may be shared with other international researchers for other studies in the future

* must provide value

● Yes ○ No

Can my samples be used in studies about population origins and ancestry?

There are researchers who are interested in studying genes to find out more about the origins and ancestry of different populations. Our genes contain information that can be used to work out where our family and our ancestors come from - our origins.

Sometimes the information from genetic studies about our ancestry may be different to what we already believe about where our family came from, or who our ancestors were, but this information can also help us understand more about human history and how people migrated around the world in the past.

We will ask you if you would like your DNA sample to be included in studies about population origins and ancestry - you do not have to agree for your sample to be used in this kind of study, it is your choice.

Do you agree for us to share your DNA sample for genetic analysis in other research studies about population origins and ancestry?

* must provide value

● Yes ○ No

Figure 2.3. The layout of the consent was changed so that consent questions came directly after the corresponding participant information. This figure also shows the simple language used and how tiered informed consent was implemented in the e-consent framework.

2.4. Utility

2.4.1. E-consent platform

One of the barriers to the uptake and implementation of e-consent is the choice of a hosting platform and the costs associated with setting it up (De Sutter et al., 2021; Haussen et al., 2017). This tiered e-consent platform was set-up in REDCap because REDCap already has an inbuilt e-consent framework which has been tested and shown to support various types of e-consent models (Chen et al., 2019; Frelich et al., 2015; Haussen et al., 2017; Lawrence et al., 2020) and is freely available on a licence agreement to organisations that are part of the REDCap consortium (Harris et al., 2019). The REDCap consortium currently comprises of more than 4000 institutions in 137 countries and membership has the added advantage of free access to technical support and improvements to the platform (Harris et al., 2019; Lawrence et al., 2020). In addition, using REDCap as the hosting platform, has the added advantage of having a single database for capturing and storing all research related data and this functionality was demonstrated for a Tuberculosis database (TBDBT) by Allie *et al.*(2020) (Allie et al., 2021).

2.4.2. Administering tiered e-consent

One of the objectives of this tiered e-consent framework was to improve participant understanding of human genomics research so that they could make truly informed consent. So, while e-consent modules are commonly designed to be participant self-administered (Doerr et al., 2016; Haas et al., 2021; Haussen et al., 2017; Kim et al., 2017; Wilbanks, 2018), the main consent document in the tiered e-consent framework will be administered face-to-face by a trained member of the research team. This mode was preferred because it affords the participant the opportunity to ask questions if they seek clarity and numerous studies have shown that participants prefer to interact with the research team as this is associated with building rapport and establishing trust (De Sutter et al., 2021; Nishimura et al., 2013; Vanaken & Masand, 2019). In addition, because this framework was developed for use in low and middle-income countries, a self-administered e-consent would not be practical. This is because REDCap is an online platform and surveys are sent to participants as a link therefore, this would potentially exclude participants who do not have access to a smart device or an internet connection and those who have limited digital literacy particularly the elderly and those in rural areas (De Sutter et al., 2020; Simon et al., 2018; Vanaken & Masand, 2019).

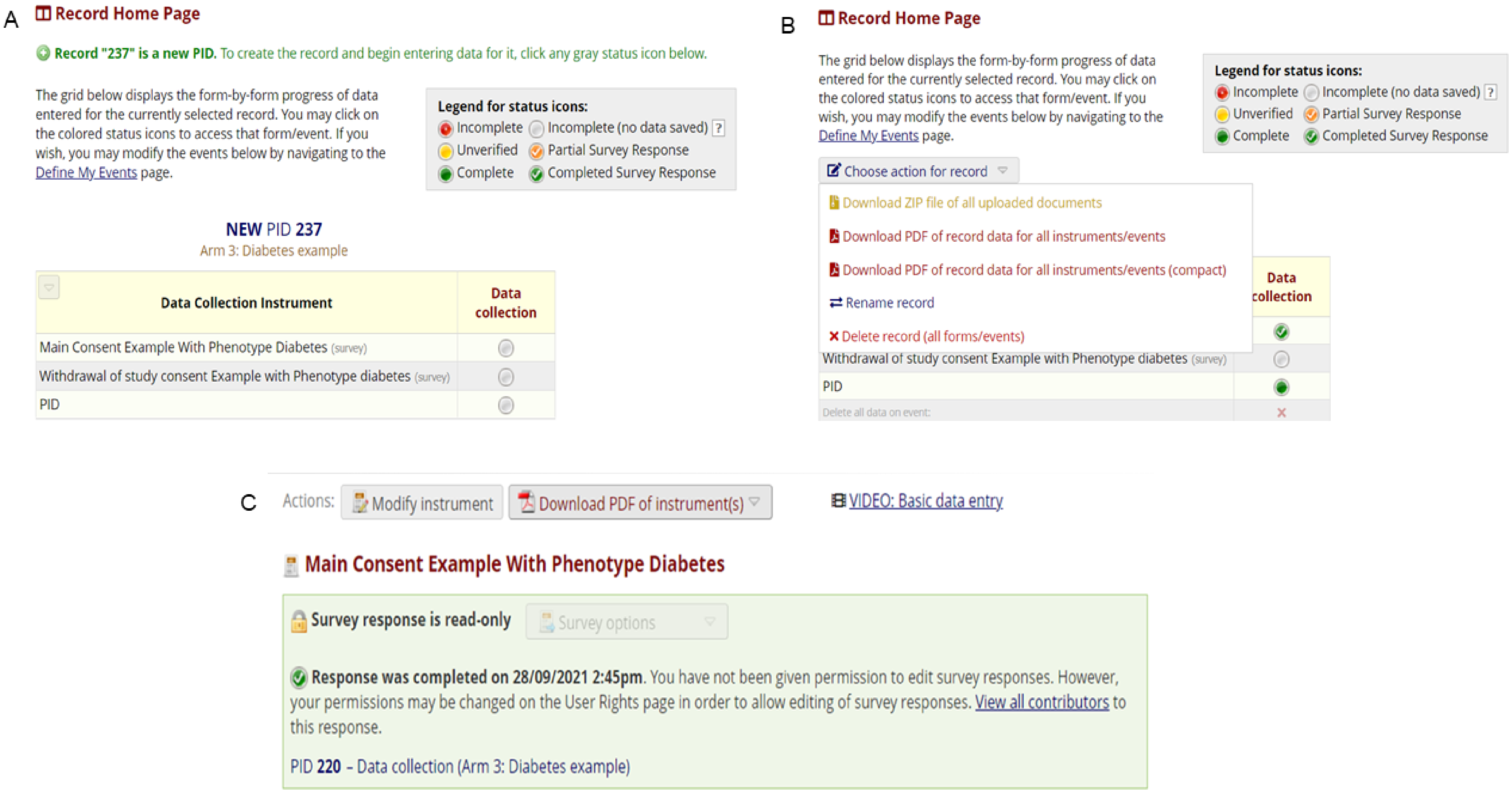
2.4.3. Data capture and storage

The main consent document uses tiered informed consent (Nembaware et al., 2019) and captures eleven different types of consent in one document (Table 2.1). The consents listed are the most common in human genomic research, but the list is not exhaustive, and users of this framework can choose which elements to include or leave out in their consent form based on their research needs. The 'Add/Edit records' function under the data collection page on the e-consent framework is used to initiate the consent process and launch the consent documents as surveys. REDCap automatically assigns a new consent survey with a unique record name (PID) which is numeric, system generated and cannot be changed (Figure 2.4A). In addition, to ensure data quality and integrity, REDCap has mandated auto numbering for all survey instruments, so that users cannot manually name new records a feature which ensures records do not share a PID.

Table 2.1. List of the type of consents that are available in the main consent of the tiered e-consent framework.

- Primary consent for collecting biospecimens and health data for specific disease in current study.
- Consent for access to medical records
- Consent for return of individual results
- Consent for return of individual results that are actionable and/or treatable
- Consent for return of individual results that are NOT actionable and/or treatable
- Consent for inclusion of individual data in genetic summary data
- Consent for use of genetic and health data for future studies on specific disease
- Consent for use of genetic and health data for future studies on other health conditions or related health processes
- Consent to re-contact for future studies
- Consent for use of genetic and health data in international studies
- Consent for use of genetic data in population origins and ancestry studies

If the consent process is not completed in one sitting, the records page has a dashboard which shows the status of each record (Figure 2.4B), and the current progress can be saved and concluded later. The PID is central to data retrieval, because once assigned it is linked to all data collection documents in the e-consent module for each participant. To retrieve an existing record, it is queried by PID and while this may be practical for a few records it will be impractical for projects with many participants. To mitigate this, a participant specific custom record label (see Supplementary file 2.4) such as the participant study ID can be appended to the PID allowing the user to retrieve individual participant records easily. The records home page also allows for records linked to a PID to be downloaded as PDFs and shared with those authorised to view them. To ensure data security and privacy, the consent documents in the tiered e-consent framework, are strictly for collecting consent information and do not collect sensitive participant information such as demographic data or contact details. In addition, to ensure data integrity, REDCap has the functionality to assign user roles and permissions for accessing, editing and/or deleting existing records after they have been verified and signed by the participant (Figure 2.4C).



Record "237" is a new PID. To create the record and begin entering data for it, click any gray status icon below.

The grid below displays the form-by-form progress of data entered for the currently selected record. You may click on the colored status icons to access that form/event. If you wish, you may modify the events below by navigating to the Define My Events page.

Legend for status icons:

🔴 Incomplete	○ Incomplete (no data saved) ?
🟡 Unverified	🟠 Partial Survey Response
🟢 Complete	✓ Completed Survey Response

NEW PID **237**
Arm 3: Diabetes example

Data Collection Instrument	Data collection
Main Consent Example With Phenotype Diabetes (survey)	○
Withdrawal of study consent Example with Phenotype diabetes (survey)	○
PID	○

The grid below displays the form-by-form progress of data entered for the currently selected record. You may click on the colored status icons to access that form/event. If you wish, you may modify the events below by navigating to the Define My Events page.

Legend for status icons:

🔴 Incomplete	○ Incomplete (no data saved) ?
🟡 Unverified	🟠 Partial Survey Response
🟢 Complete	✓ Completed Survey Response

Choose action for record
Download ZIP file of all uploaded documents
Download PDF of record data for all instruments/events
Download PDF of record data for all instruments/events (compact)
Rename record
Delete record (all forms/events)

	Data collection
	✓
Withdrawal of study consent Example with Phenotype diabetes (survey)	○
PID	●
Delete all data on event:	×

C
Actions:
Modify instrument
Download PDF of instrument(s)
VIDEO: Basic data entry
Main Consent Example With Phenotype Diabetes
Survey response is read-only
Survey options

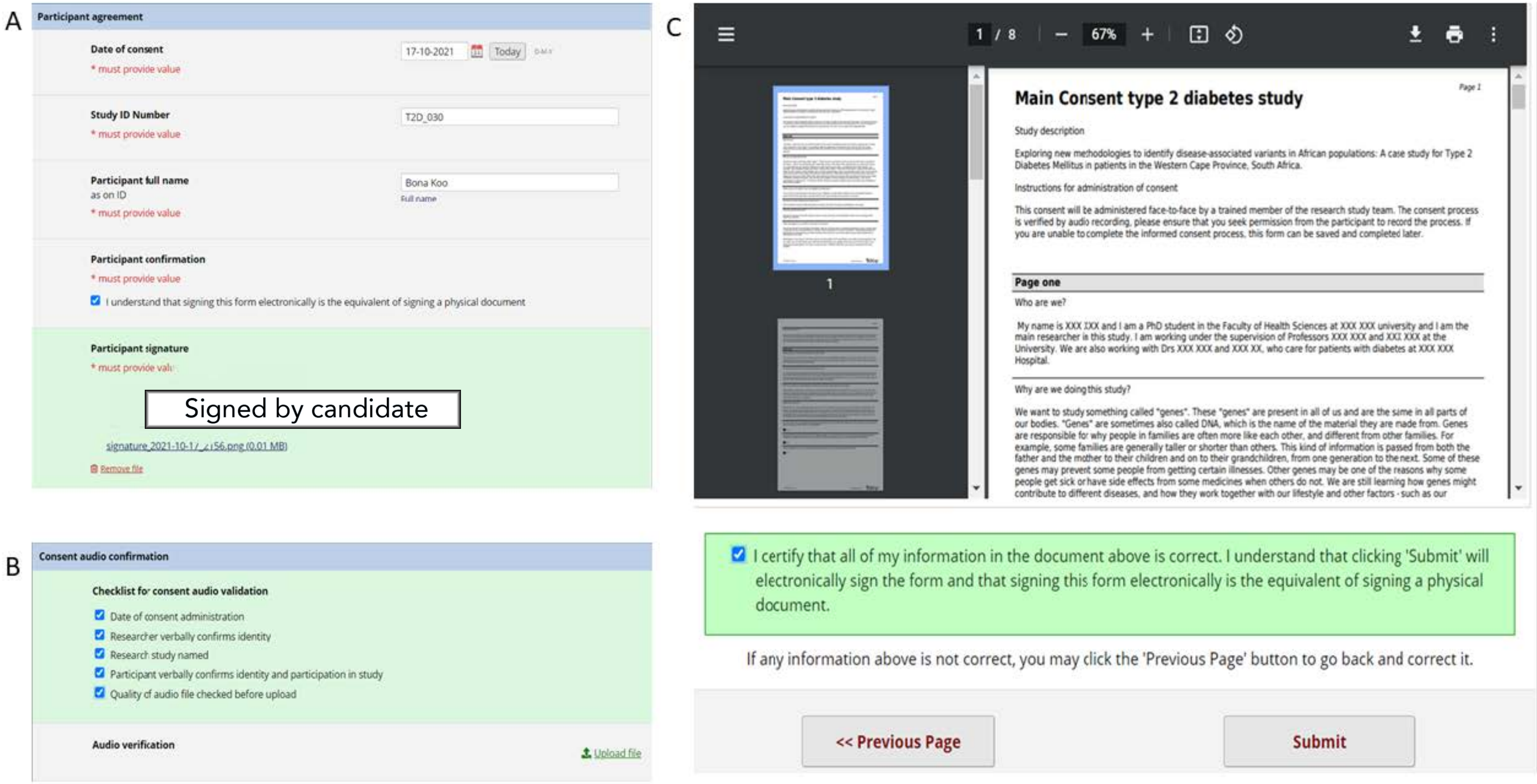
✓ **Response was completed on 28/09/2021 2:45pm**. You have not been given permission to edit survey responses. However, your permissions may be changed on the User Rights page in order to allow editing of survey responses. View all contributors to this response.

PID **220** – Data collection (Arm 3: Diabetes example)

Figure 2.4. An overview of the records home page in REDCap. (A) Assignment of a PID to initiate collection of consent data, (B) dashboard showing the status of data collection in each instrument and (C) an example of user permission assignment

2.4.4. Data verification

To meet legal and regulatory requirements all consent documents are validated by date-stamped electronic signature (Figure 2.5A). In cases where an electronic signature is not legally recognised, the consent can still be done online, and the form downloaded and signed by hand. The signed form or the signature itself can then be scanned and uploaded as an attachment alongside the signature field in the e-consent form. In this instance we would recommend archiving the signed paper forms in case they are required in the future. This will ensure that all consent data is still captured electronically directly into the REDCap database. For transparency, the e-consenting process will have two verification steps. The first is audio verification (Figure 2.5B) for which participant permission will be sough before the consenting begins, and the audio file generated can also be uploaded and stored in REDCap. The second is through the 'auto-archiver' feature (Figure 2.5C) which gives the participant an opportunity to verify that their choices were captured accurately before the consent is finalised.



Participant agreement

Date of consent 17-10-2021 Today D-M-Y
* must provide value

Study ID Number T2D_030
* must provide value

Participant full name
as on ID
* must provide value
Bona Koo
Full name

Participant confirmation
* must provide value
☑ I understand that signing this form electronically is the equivalent of signing a physical document

Participant signature
* must provide valı

Signed by candidate

signature_2021-10-1/_¿ı56.png (0.01 MB)

Remove file

Consent audio confirmation

Checklist for consent audio validation

☑ Date of consent administration
☑ Researcher verbally confirms identity
☑ Research study named
☑ Participant verbally confirms identity and participation in study
☑ Quality of audio file checked before upload

Audio verification Upload file

1 / 8 67%

1

Page 1

Main Consent type 2 diabetes study

Study description

Exploring new methodologies to identify disease-associated variants in African populations: A case study for Type 2 Diabetes Mellitus in patients in the Western Cape Province, South Africa.

Instructions for administration of consent

This consent will be administered face-to-face by a trained member of the research study team. The consent process is verified by audio recording, please ensure that you seek permission from the participant to record the process. If you are unable to complete the informed consent process, this form can be saved and completed later.

Page one

Who are we?

My name is XXX IXX and I am a PhD student in the Faculty of Health Sciences at XXX XXX university and I am the main researcher in this study. I am working under the supervision of Professors XXX XXX and XXI XXX at the University. We are also working with Drs XXX XXX and XXX XX, who care for patients with diabetes at XXX XXX Hospital.

Why are we doing this study?

We want to study something called "genes". These "genes" are present in all of us and are the sime in all parts of our bodies. "Genes" are sometimes also called DNA, which is the name of the material they are made from. Genes are responsible for why people in families are often more like each other, and different from other families. For example, some families are generally taller or shorter than others. This kind of information is passed from both the father and the mother to their children and on to their grandchildren, from one generation to the next. Some of these genes may prevent some people from getting certain illnesses. Other genes may be one of the reasons why some people get sick or have side effects from some medicines when others do not. We are still learning how genes might contribute to different diseases, and how they work together with our lifestyle and other factors - such as our

☑ I certify that all of my information in the document above is correct. I understand that clicking 'Submit' will electronically sign the form and that signing this form electronically is the equivalent of signing a physical document.

If any information above is not correct, you may click the 'Previous Page' button to go back and correct it.

<< Previous Page Submit

Figure 2.5. Verification and validation of the e-consent process by (A) electronic signature, (B) audio verification and (C) review and certification of the consent choices made by the participant before form submission.

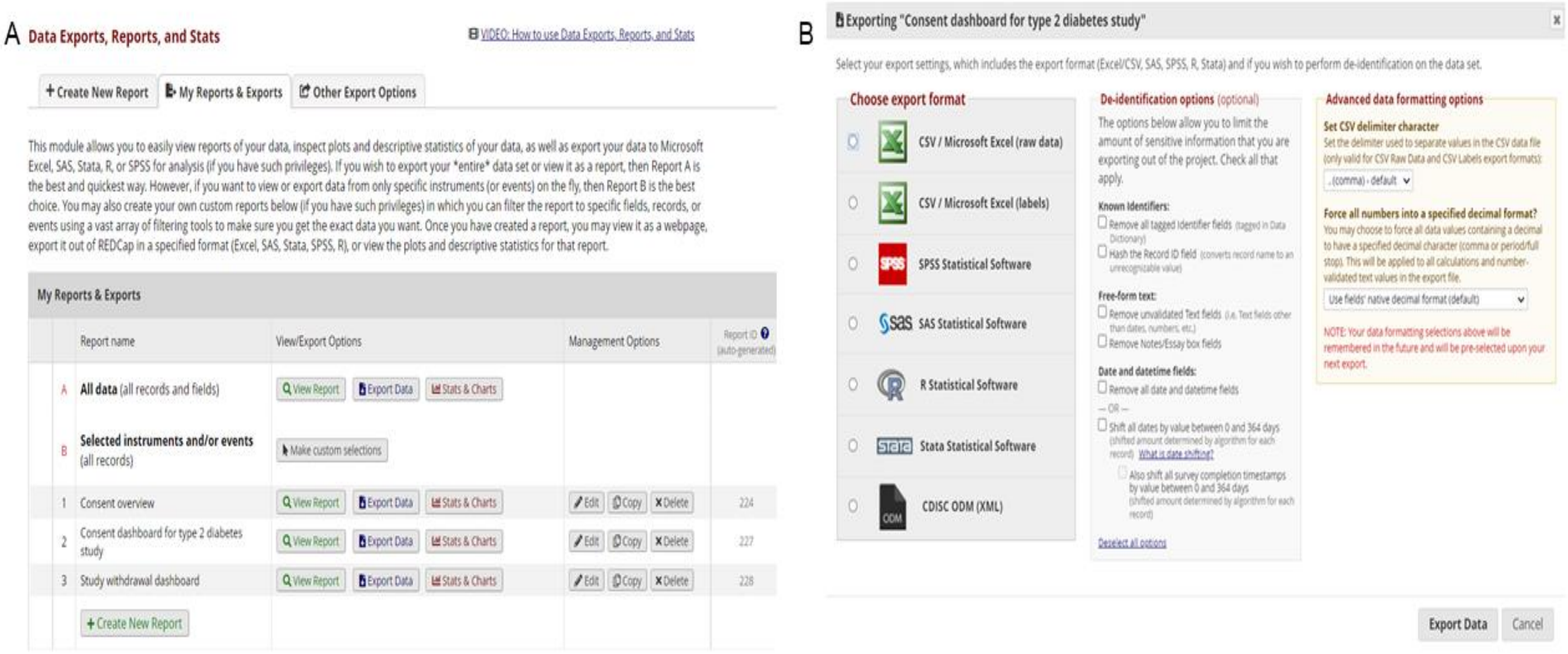
2.4.5. Data query and export

All data that are entered into the tiered e-consent module are automatically stored in REDCap and can be viewed and downloaded from the Reports tab (Figure 2.6A). REDCap automatically creates reports, but also allows for the customisation of reports to suite specific research needs by allowing users to select which data elements to include in each report. For the tiered e-consent module, we created two customised reports, being the consent dashboard (see Supplementary file 2.4) and the study withdrawal dashboard (see Supplementary file 2.5), which contain information on who has consented and/ or withdrawn from the study both at an individual level and for the entire study population. In addition, because this is tiered consent, the study population data is summarised for each type of consent covered in the main consent (see Supplementary file 2.6).

REDCap supports automated export of study reports, and the data can be downloaded in a format suitable to a selection of commonly used statistical packages (Figure 2.6B). These reports also allow researchers to easily monitor the progress of their recruitment process in real time for in-house use and for submitting study progress reports to institutional ethics review boards. To protect participant privacy, there is an option to hide all tagged identifier fields and/or hash-to the record ID field. In addition, because the 'data exports, reports and stats' feature make it easy for researchers to query the database, identify consenting individuals and download their consent data this will facilitate ease of collaboration among researchers conducting human genomic studies.

2.4.6. Withdrawal of consent

An important feature of voluntary participation in research, is that participants can withdraw from the study whenever they wish. To accurately document participants who wish to withdraw their consent, the withdrawal of consent document is used. The consent can be partial, or complete and 'branching logic' (Figure 2.7) is used to differentiate the two. If a participant selects to withdraw from the study completely (Figure 2.7A) then the next option is to provide a reason and then sign. However, if a participant wishes to withdraw only certain parts of their consent, then a list of options opens, and the relevant ones are selected. Following this, the rest of the process is the same as for complete withdrawal.


A
Data Exports, Reports, and Stats
VIDEO: How to use Data Exports, Reports, and Stats
+ Create New Report
My Reports & Exports
Other Export Options

This module allows you to easily view reports of your data, inspect plots and descriptive statistics of your data, as well as export your data to Microsoft Excel, SAS, Stata, R, or SPSS for analysis (if you have such privileges). If you wish to export your *entire* data set or view it as a report, then Report A is the best and quickest way. However, if you want to view or export data from only specific instruments (or events) on the fly, then Report B is the best choice. You may also create your own custom reports below (if you have such privileges) in which you can filter the report to specific fields, records, or events using a vast array of filtering tools to make sure you get the exact data you want. Once you have created a report, you may view it as a webpage, export it out of REDCap in a specified format (Excel, SAS, Stata, SPSS, R), or view the plots and descriptive statistics for that report.

My Reports & Exports

	Report name	View/Export Options	Management Options	Report ID (auto-generated)
A	**All data** (all records and fields)	View Report Export Data Stats & Charts		
B	**Selected instruments and/or events** (all records)	Make custom selections		
1	Consent overview	View Report Export Data Stats & Charts	Edit Copy Delete	224
2	Consent dashboard for type 2 diabetes study	View Report Export Data Stats & Charts	Edit Copy Delete	227
3	Study withdrawal dashboard	View Report Export Data Stats & Charts	Edit Copy Delete	228
	+ Create New Report			

Exporting "Consent dashboard for type 2 diabetes study"

Select your export settings, which includes the export format (Excel/CSV, SAS, SPSS, R, Stata) and if you wish to perform de-identification on the data set.

Choose export format

- ○ CSV / Microsoft Excel (raw data)
- ○ CSV / Microsoft Excel (labels)
- ○ SPSS Statistical Software
- ○ SAS Statistical Software
- ○ R Statistical Software
- ○ Stata Statistical Software
- ○ CDISC ODM (XML)

De-identification options (optional)

The options below allow you to limit the amount of sensitive information that you are exporting out of the project. Check all that apply.

Known identifiers:

□ Remove all tagged Identifier fields (tagged in Data Dictionary)

□ Hash the Record ID field (converts record name to an unrecognizable value)

Free-form text:

□ Remove unvalidated Text fields (i.e. Text fields other than dates, numbers, etc.)

□ Remove Notes/Essay box fields

Date and datetime fields:

□ Remove all date and datetime fields

-- OR --

□ Shift all dates by value between 0 and 364 days (shifted amount determined by algorithm for each record) What is date shifting?

□ Also shift all survey completion timestamps by value between 0 and 364 days (shifted amount determined by algorithm for each record)

Deselect all options

Advanced data formatting options

Set CSV delimiter character

Set the delimiter used to separate values in the CSV data file (only valid for CSV Raw Data and CSV Labels export formats):

, (comma) - default

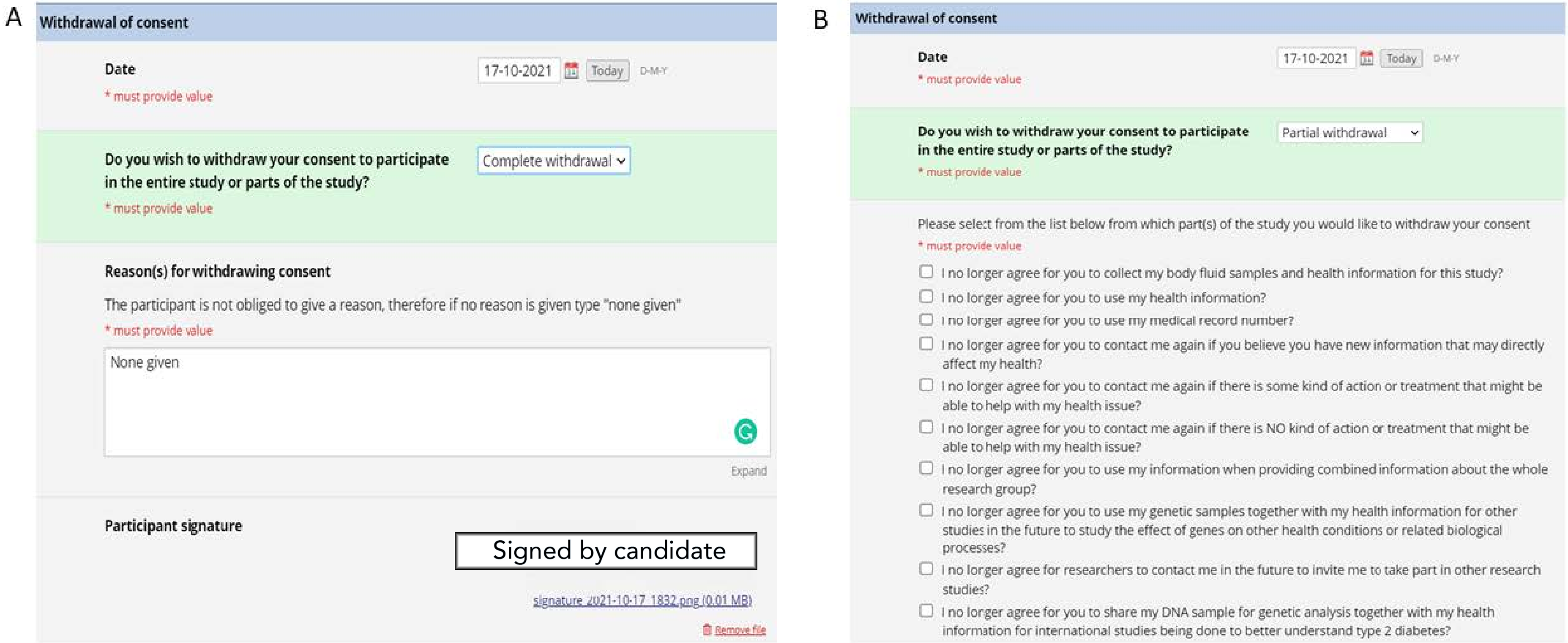
Force all numbers into a specified decimal format?

You may choose to force all data values containing a decimal to have a specified decimal character (comma or period/full stop). This will be applied to all calculations and number-validated text values in the export file.

Use fields' native decimal format (default)

NOTE: Your data formatting selections above will be remembered in the future and will be pre-selected upon your next export.

Figure 2.6. The data export, reports, and statistics page in REDCap. (A) An overview of all reports in the database and (B) the dashboard for the automated export of data from REDCap.



Withdrawal of consent

Date

* must provide value

17-10-2021 Today D-M-Y

Do you wish to withdraw your consent to participate in the entire study or parts of the study?

* must provide value

Complete withdrawal

Reason(s) for withdrawing consent

The participant is not obliged to give a reason, therefore if no reason is given type "none given"

* must provide value

None given

Expand

Participant signature

Signed by candidate

signature 2021-10-17 1832.png (0.01 MB)

Remove file

Withdrawal of consent

Date

* must provide value

17-10-2021 Today D-M-Y

Do you wish to withdraw your consent to participate in the entire study or parts of the study?

* must provide value

Partial withdrawal

Please select from the list below from which part(s) of the study you would like to withdraw your consent

* must provide value

- □ I no longer agree for you to collect my body fluid samples and health information for this study?
- □ I no longer agree for you to use my health information?
- □ I no longer agree for you to use my medical record number?
- □ I no longer agree for you to contact me again if you believe you have new information that may directly affect my health?
- □ I no longer agree for you to contact me again if there is some kind of action or treatment that might be able to help with my health issue?
- □ I no longer agree for you to contact me again if there is NO kind of action or treatment that might be able to help with my health issue?
- □ I no longer agree for you to use my information when providing combined information about the whole research group?
- □ I no longer agree for you to use my genetic samples together with my health information for other studies in the future to study the effect of genes on other health conditions or related biological processes?
- □ I no longer agree for researchers to contact me in the future to invite me to take part in other research studies?
- □ I no longer agree for you to share my DNA sample for genetic analysis together with my health information for international studies being done to better understand type 2 diabetes?

Figure 2.7. Implementation of 'branching logic' in the withdrawal of consent document. (A) Complete withdrawal and (B) partial withdrawal with the option for the participant to select which elements they want to withdraw their consent

2.5. Discussion

Part of ethical research is ensuring that we make the best use of collected data and specimens, in line with the permissions that are given by participants. The use of broad consent has created some barriers to onward use of data, as it is not always clear exactly what participants have or have not agreed to; and it also makes it difficult to respect the individual preferences and autonomy of participants (Tiffin, 2018). With the advent of legislation that protects privacy of individuals, like the General Data Protection Regulation (GDPR) in the European Union (EU), or the Protection of Personal Information (POPI) Act in South Africa, it is important to have consent from individuals specifically for sharing their health and data with other researchers and/or across international borders. While asking consent for each specific use might limit re-use for new types of research in the future that we do not yet know about, including consent specifically to be re-contacted means that researchers can contact participants about new types of studies in the future. Whilst not all participants might agree to re-contact for future studies, for those that do, this provides an option to consult them directly without the researcher or an ethics review board making these important decisions on behalf of the participant but without their knowledge.

Building informed consent processes without prior experience can be daunting, so we have aimed to assist researchers by developing this template that reminds researchers of the types of consent they can request for genomics studies and assists them with suggestions for the language they might use for participant information and consent questions; whilst allowing them the freedom to include or exclude certain modules and modify the language that they use. Finally, immediate electronic capture of the consents given by participants can facilitate accurate and efficient onward sharing of data and samples according to participant preferences that can be easily electronically queried. This can replace the current common practice of unwieldy storage of paper consent forms that need to be reviewed individually to determine which data or samples can be re-used. The use of this e-consent module can thus facilitate efficient and ethical data- and sample-sharing, whilst respecting the specific preferences and choices of each participant.

Whilst this REDCap template utilises a digital approach to presenting and capturing the informed consent process, which comes with the described advantages such as improved data fidelity and streamlined databasing of participant choices, the fundamental process and material content of tiered informed consent remains consistent with current paper-based tiered informed consent processes for health genomics research (Nembaware et al., 2019, 2020). This point can be communicated clearly to ethics review committees assessing the use of the template for the first time. As with current practice, and as described here, it remains important to validate the informed consent process to ensure it is locally relevant, through community engagement, for example by holding community-based focus groups to evaluate local accessibility of the content. Other important inputs include training researchers in the use of the digital informed consent process to ensure high quality data collection as well as to ensure participants understand how the digital tool is being used. Ongoing data quality control can also ensure effective use and appropriate data capture with the REDCap informed consent tool. Through these approaches, participants, researchers, and ethics review boards can gain confidence that the informed consent process is operating as intended.

2.6. References

Allie, T., Jackson, A., Ambler, J., Johnston, K., Bruyn, E. D., Schultz, C., Boloko, L., Wasserman, S., Davis, A., Meintjes, G., Wilkinson, R. J., & Tiffin, N. (2021). TBDBT: A TB DataBase Template for collection of harmonized TB clinical research data in REDCap, facilitating data standardisation for inter-study comparison and meta-analyses. *PLOS ONE*, *16*(3), e0249165. https://doi.org/10.1371/journal.pone.0249165

Artal, R., & Rubenfeld, S. (2017). Ethical issues in research. *Best Practice & Research Clinical Obstetrics & Gynaecology*, *43*, 107–114. https://doi.org/10.1016/j.bpobgyn.2016.12.006

Chalil Madathil, K., Koikkara, R., Obeid, J., Greenstein, J. S., Sanderson, I. C., Fryar, K., Moskowitz, J., & Gramopadhye, A. K. (2013). An investigation of the efficacy of electronic consenting interfaces of research permissions management system in a hospital setting. *International Journal of Medical Informatics*, *82*(9), 854–863. https://doi.org/10.1016/j.ijmedinf.2013.04.008

Chen, C., Turner, S. P., Sholle, E. T., Brown, S. W., Blau, V. L. I., Brouwer, J. P., Lewis, A. N., Cole, C. L., Nanus, D. M., Shah, M. A., Leonard, J. P., & Campion, T. R. (2019). Evaluation of a REDCap-based Workflow for Supporting Federal Guidance for

Electronic Informed Consent. *AMIA Summits on Translational Science Proceedings*, *2019*, 163–172.

Dankar, F. K., Gergely, M., & Dankar, S. K. (2019). Informed Consent in Biomedical Research. *Computational and Structural Biotechnology Journal*, *17*, 463–474. https://doi.org/10.1016/j.csbj.2019.03.010

De Sutter, E., Borry, P., Geerts, D., & Huys, I. (2021). Personalized and long-term electronic informed consent in clinical research: Stakeholder views. *BMC Medical Ethics*, *22*(1), 108. https://doi.org/10.1186/s12910-021-00675-7

De Sutter, E., Zaçe, D., Boccia, S., Di Pietro, M. L., Geerts, D., Borry, P., & Huys, I. (2020). Implementation of Electronic Informed Consent in Biomedical Research and Stakeholders' Perspectives: Systematic Review. *Journal of Medical Internet Research*, *22*(10), e19129. https://doi.org/10.2196/19129

Doerr, M., Suver, C., & Wilbanks, J. (2016). *Developing a Transparent, Participant-Navigated Electronic Informed Consent for Mobile-Mediated Research* (SSRN Scholarly Paper ID 2769129). Social Science Research Network. https://doi.org/10.2139/ssrn.2769129

Frelich, M. J., Bosler, M. E., & Gould, J. C. (2015). Research Electronic Data Capture (REDCap) electronic Informed Consent Form (eICF) is compliant and feasible in a clinical research setting. *International Journal of Clinical Trials*, *2*(3), 51–55. https://doi.org/10.18203/2349-3259.ijct20150591

Haas, M. A., Teare, H., Prictor, M., Ceregra, G., Vidgen, M. E., Bunker, D., Kaye, J., & Boughtwood, T. (2021). 'CTRL': An online, Dynamic Consent and participant engagement platform working towards solving the complexities of consent in genomic research. *European Journal of Human Genetics*, *29*(4), 687–698. https://doi.org/10.1038/s41431-020-00782-w

Harris, P. A., Taylor, R., Minor, B. L., Elliott, V., Fernandez, M., O'Neal, L., McLeod, L., Delacqua, G., Delacqua, F., Kirby, J., & Duda, S. N. (2019). The REDCap consortium: Building an international community of software platform partners.

Journal of Biomedical Informatics, *95*, 103208. https://doi.org/10.1016/j.jbi.2019.103208

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, *42*(2), 377–381. https://doi.org/10.1016/j.jbi.2008.08.010

Haussen, D. C., Doppelheuer, S., Schindler, K., Grossberg, J. A., Bouslama, M., Schultz, M., Perez, H., Hall, A., Frankel, M., & Nogueira, R. G. (2017). Utilization of a Smartphone Platform for Electronic Informed Consent in Acute Stroke Trials. *Stroke*, *48*(11), 3156–3160. https://doi.org/10.1161/STROKEAHA.117.018380

Kim, H., Bell, E., Kim, J., Sitapati, A., Ramsdell, J., Farcas, C., Friedman, D., Feupe, S. F., & Ohno-Machado, L. (2017). iCONCUR: Informed consent for clinical data and bio-sample use for research. *Journal of the American Medical Informatics Association: JAMIA*, *24*(2), 380–387. https://doi.org/10.1093/jamia/ocw115

Lawrence, C. E., Dunkel, L., McEver, M., Israel, T., Taylor, R., Chiriboga, G., Goins, K. V., Rahn, E. J., Mudano, A. S., Roberson, E. D., Chambless, C., Wadley, V. G., Danila, M. I., Fischer, M. A., Joosten, Y., Saag, K. G., Allison, J. J., Lemon, S. C., & Harris, P. A. (2020). A REDCap-based model for electronic consent (eConsent): Moving toward a more personalized consent. *Journal of Clinical and Translational Science*, *4*(4), 345–353. https://doi.org/10.1017/cts.2020.30

Lentz, J., Kennett, M., Perlmutter, J., & Forrest, A. (2016). Paving the way to a more effective informed consent process: Recommendations from the Clinical Trials Transformation Initiative. *Contemporary Clinical Trials*, *49*, 65–69. https://doi.org/10.1016/j.cct.2016.06.005

Nembaware, V., Johnston, K., Diallo, A. A., Kotze, M. J., Matimba, A., Moodley, K., Tangwa, G. B., Torrorey-Sawe, R., & Tiffin, N. (2019). A framework for tiered informed consent for health genomic research in Africa. *Nature Genetics*, *51*(11), 1566–1571. https://doi.org/10.1038/s41588-019-0520-x

Nembaware, V., Munung, N. S., Matimba, A., & Tiffin, N. (2020). Patient-centric research in the time of COVID-19: Conducting ethical COVID-19 research in Africa. *BMJ Global Health*, *5*(8), e003035. https://doi.org/10.1136/bmjgh-2020-003035

Nishimura, A., Carey, J., Erwin, P. J., Tilburt, J. C., Murad, M. H., & McCormick, J. B. (2013). Improving understanding in the research informed consent process: A systematic review of 54 interventions tested in randomized control trials. *BMC Medical Ethics*, *14*(1), 28. https://doi.org/10.1186/1472-6939-14-28

Pranati. (2010). Informed consent: Are we doing enough? *Perspectives in Clinical Research*, *1*(4), 124–127. https://doi.org/10.4103/2229-3485.71769

Simon, C. M., Schartz, H. A., Rosenthal, G. E., Eisenstein, E. L., & Klein, D. W. (2018). Perspectives on Electronic Informed Consent From Patients Underrepresented in Research in the United States: A Focus Group Study. *Journal of Empirical Research on Human Research Ethics*, *13*(4), 338–348. https://doi.org/10.1177/1556264618773883

Tiffin, N. (2018). Tiered informed consent: Respecting autonomy, agency and individuality in Africa. *BMJ Global Health*, *3*(6), e001249. https://doi.org/10.1136/bmjgh-2018-001249

Trinidad, S. B., Fullerton, S. M., Bares, J. M., Jarvik, G. P., Larson, E. B., & Burke, W. (2012). Informed Consent in Genome-Scale Research: What Do Prospective Participants Think? *AJOB Primary Research*, *3*(3), 3–11. https://doi.org/10.1080/21507716.2012.662575

Vanaken, H., & Masand, S. N. (2019). Awareness and Collaboration Across Stakeholder Groups Important for eConsent Achieving Value-Driven Adoption. *Therapeutic Innovation & Regulatory Science*, *53*(6), 724–735. https://doi.org/10.1177/2168479019861924

Wilbanks, J. (2018). Design Issues in E-Consent. *Journal of Law, Medicine & Ethics*, *46*(1), 110–118. https://doi.org/10.1177/1073110518766025

3. Chapter 3: Diabetes in an TB and HIV-endemic South African population: Analysis of a virtual cohort using routine health data.

Tamuhla T, Dave JA, Raubenheimer P, Tiffin N. Diabetes in a TB and HIV-endemic South African population: Analysis of a virtual cohort using routine health data. PLOS ONE. 2021;16: e0251303. doi:10.1371/journal.pone.0251303

Author contributions

TT - Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing.
JD & PJR - Formal analysis, Investigation, Methodology, Writing – review & editing.
NT - Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing – review & editing.

Relevance of paper to the thesis

Since the virtual genotyped cohort will be piloted in people with type 2 diabetes, the primary aim of this epidemiologic analysis was to describe the population from which these participants would be recruited. This is especially relevant here in South Africa where we have high infectious disease comorbidities. This was done by generating summary statistics from the demographic, laboratory and pharmacy data using an HIV comorbidities cohort data set from the provincial health data centre. This analysis also addressed objective 2 where we explored statistical methods for modelling longitudinal data so that it can be used as a GWAS phenotype.

3.1. Abstract

Background

It is widely accepted that people living with diabetes (PLWD) are at increased risk of infectious disease, yet there is a paucity of epidemiology studies on the relationship between diabetes and infectious disease in SSA. In a region with a high burden of infectious disease, this has serious consequences for PLWD.

Methods and Findings

Using routinely collected longitudinal health data, we describe the epidemiology of diabetes in a large virtual cohort of PLWD who have a high burden of HIV and TB, from the Khayelitsha subdistrict in the Western Cape Province in South Africa. We described the relationship between previous TB, newly diagnosed TB disease and HIV infection on diabetes using HbA1c results as an outcome measure. The study population was predominately female (67%), 13% had a history of active TB disease and 18% were HIV positive. The HIV positive group had diabetes ascertained at a significantly younger age (46 years c.f. 53 years respectively, $p<0.001$) and in general had increased HbA1c values over time after their HIV diagnosis, when compared to the HIV-negative group. There was no evidence of TB disease influencing the trajectory of glycaemic control in the long term, but diabetes patients who developed active TB had higher mortality than those without TB (12.4% vs 6.7% p-value < 0.001). HIV and diabetes are both chronic diseases whose long-term management includes drug therapy, however, only 52.8% of the study population with an HIV-diabetes comorbidity had a record of diabetes treatment. In addition, the data suggest overall poor glycaemic control in the study population with only 24.5% of the participants having an HbA1c <7 % at baseline despite 85% of the study population being on diabetes treatment.

Conclusion

The epidemiologic findings in this exploratory study highlight the need for further research into diabetes outcomes in a high TB and HIV burden setting and demonstrate that routine health data are a valuable resource for understanding disease epidemiology in the general population.

3.2. Introduction

Sub-Saharan Africa (SSA) is currently undergoing an epidemiologic shift and the health systems in the region are dealing with the dual burden of infectious diseases and an increasing prevalence of non-communicable diseases (NCDs) (Africa, n.d.; Levitt, 2008) . NCDs are overtaking infectious disease as the leading cause of disability and mortality in the region (Africa, n.d.; Gouda et al., 2019; *Noncommunicable Diseases*, n.d.). This epidemiologic transition is already evident in South Africa where, although Tuberculosis (TB) was still the overall leading cause of natural deaths from 2015 - 2017, in the same time period, Diabetes Mellitus (DM) was the second leading cause of death (Africa, n.d.). The burden of DM is putting a strain on already struggling public health systems, and with an estimated 19 million people with diabetes in the region currently, projected to increase to 29 million by 2030, SSA is facing an impending diabetes epidemic (*IDF Diabetes Atlas 9th Edition 2019*, n.d.)

More than 90% of diabetes in SSA is type 2 diabetes mellitus (T2DM) (Hall et al., 2011) which is thought to be largely fuelled by lifestyle changes brought about by a surge in rural-urban migration (Bertram et al., 2013). The diabetes epidemic in SSA including South Africa is further complicated by the ongoing HIV epidemic. South Africa is already implementing the UNAIDS 90-90-90 strategy which aims to get 90% of all those who test positive for HIV on anti-retroviral therapy (ART) (Williams et al., 2017). This widespread use of ART has significantly increased the life expectancy of people living with HIV (PLWH), and the country is now supporting an aging HIV population that are developing comorbidities such as DM associated with aging which might also occur at earlier ages than in the general population (Rasmussen et al., 2015; Schouten et al., 2014). Studies have shown that in addition to demographic and lifestyle risk factors for DM the chronic use of ART, especially HIV protease inhibitors (PIs) and non-nucleoside reverse transcriptase inhibitors (NNRTIs), also contribute to the risk of developing DM (Araujo et al., 2014; Dave et al., 2011; Levitt et al., 2016; Nansseu et al., 2018).

While the widespread use of ART is reducing HIV/AIDS related morbidity and mortality - especially due to TB co-infection which is the leading cause of death in HIV positive people - it could potentially fuel the resurgence of diabetes-associated TB (Harries et al., 2011, 2015; Reid et al., 2013; Williams et al., 2017). The relationship between TB and DM is well established (Restrepo, 2007) and studies have shown that diabetes increases the risk of developing active TB, recurrent TB and severe TB disease and results in worse TB treatment outcomes (Cheng et al., 2017; Chiang et al., 2015; Leung et al., 2008; Munseri et al., 2019; Pizzol et al., 2017, 2018; Workneh et al., 2017). The threat of a TB-DM dual epidemic in South Africa is a cause for concern given that the country is in the top eight highest TB burden countries, and in 2019 accounted for 3.6% of the global total of people who developed active TB (*Global Tuberculosis Report 2020*, n.d.), and the trilateral overlap with HIV may therefore have implications for TB control (Oni et al., 2017). In addition, most of the DM in SSA including South African is undiagnosed until it presents with severe symptoms, so by the time most people get diagnosed they are already at risk of DM-related complications (Mbanya et al., 2010; Stokes et al., 2017).

It is widely accepted that people living with diabetes (PLWD) are at increased risk of infectious disease, yet there is a paucity of epidemiology studies on the relationship between diabetes and infectious disease in SSA. In a region with a heavy burden of infectious disease, this has serious consequences for PLWD in the region. Here, we describe the epidemiology of diabetes in a large virtual cohort of PLWD from the Western Cape Province in South Africa, who have a high burden of HIV and TB, using routinely collected longitudinal health data. We describe the relationship between previous and newly diagnosed TB disease and HIV infection and pre-existing diabetes using National Glycohemoglobin Standardization Program (NGSP) HbA1c as an outcome measure.

3.3. Methods

3.3.1. Ethics

Ethics approval was granted by the University of Cape Town (HREC REF: 509/2019) and data access was approved by Western Cape Government Health (WCGH), South Africa. All data were de-identified and data perturbation was employed by the Provincial Health Data Centre (PHDC, WCGH) prior to release, so that the data used were anonymised and cannot be reidentified. Data transfer was effected through secure platforms using AES256 encryption and password protection, and analysis was undertaken on a secured, firewall-protected server. Re-use of this dataset requires approval from the PHDC, and the authors can be contacted to advise on this process.

3.3.2. Study population

The study population was selected from the Western Cape Population as represented in the PHDC, a health information exchange containing routine health data for about 7 million healthcare clients, collated daily from multiple electronic health data sources in the Western Cape Province, South Africa (Boulle et al., 2019). Inclusion criteria were: (1) Having attended at least one Government Health Facility in the Khayelitsha sub-district in the Western Cape, South Africa, in the period 1 January 2016 to 31 December 2017, (2) aged 18 or older by December 2017 and (3) diagnosis of diabetes was inferred in people who ever had an HBA1c of ≥ 6.5% (“Use of Glycated Haemoglobin (HbA1c) in the Diagnosis of Diabetes Mellitus,” 2011), a 2-hr glucose of ≥ 11.1mmol/l after an oral glucose tolerance test, had been dispensed oral hypoglycaemic agents used exclusively for the management of diabetes or insulin. The Khayelitsha subdistrict is a high-density urban area with large areas of informal housing and generally poor socioeconomic conditions. Exclusion criteria were: Diabetes ascertainment at less than 18 years of age, used as a proxy for early onset Type 1 Diabetes; and diabetes ascertainment occurring during pregnancy, used as a proxy for gestational diabetes (Figure 3.1). HbA1c is used as a diabetes outcome measure because it is the gold standard for diagnosing and monitoring diabetes control (“International Expert Committee Report on the Role of the A1C Assay in the Diagnosis of Diabetes,” 2009; “Use of Glycated Haemoglobin (HbA1c) in the Diagnosis of Diabetes Mellitus,” 2011), so people who did not have any recorded HbA1c values were also excluded from the analysis. HbA1c may underestimate glycaemia in PLWH but despite this remains highly specific for DM diagnosis (Eckhardt et al., 2012). When analysing medications, a subset was created which excluded those with no diabetes treatment records (Figure 3.1).

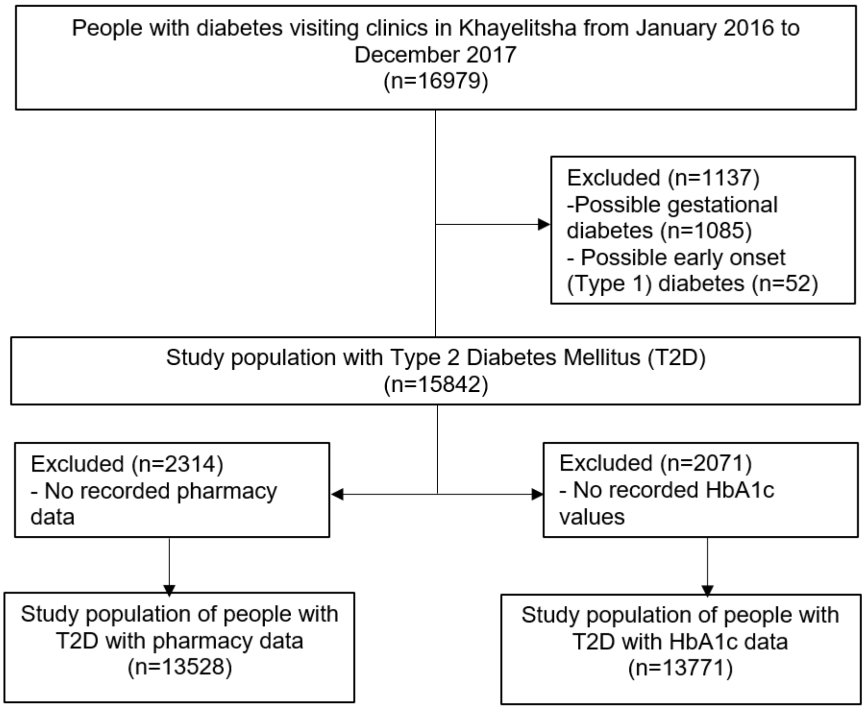

People with diabetes visiting clinics in Khayelitsha from January 2016 to December 2017 (n=16979)
Excluded (n=1137) -Possible gestational diabetes (n=1085) - Possible early onset (Type 1) diabetes (n=52)
Study population with Type 2 Diabetes Mellitus (T2D) (n=15842)
Excluded (n=2314) - No recorded pharmacy data
Excluded (n=2071) - No recorded HbA1c values
Study population of people with T2D with pharmacy data (n=13528)
Study population of people with T2D with HbA1c data (n=13771)

Figure 3.1. Flow chart showing the selection of the study population from the PHDC routine health data.

Retrospective PHDC data for 13 771 individuals with recorded HbA1c values were analysed together with population demographics (data as of 31 December 2017) using descriptive statistics. A diagnosis is inferred by the PHDC using laboratory and pharmacy data, and is not a clinical diagnosis, so is referred to as ‘ascertainment’ to make this distinction, and is described in (Boulle et al., 2019). Diabetes metrics include: Age at ascertainment, ‘Linkage to HbA1c testing’ was defined as having had a recorded HbA1c laboratory test result within one year of the last recorded diabetes-related health facility encounter, ‘Ever started diabetes treatment’ was defined as those with a recorded diabetes treatment start date and ‘linkage to diabetes treatment’ as those who had a record of diabetes drugs being prescribed within one year of their last recorded diabetes-related health facility encounter. ‘Baseline HbA1c’ was defined as the first recorded HbA1c value either at diabetes ascertainment or within the first year after diabetes ascertainment. TB metrics include: date of ascertainment of TB episodes for individuals from PHDC inferred episodes, ‘Ever had Tuberculosis’ was defined as having had a TB episode at any time in an individual’s recorded medical history and ‘TB-Diabetes comorbidity’ as having a recorded TB episode after Diabetes ascertainment. HIV metrics include the date of ascertainment of HIV from PHDC inferred episode data. HIV status was determined using date of HIV ascertainment and a record of initiation on HIV anti-retroviral therapy (ART).

Summary statistics were calculated for the study population. For continuous data, median and interquartile range were calculated and for grouped data, percentages were calculated. For median values, the Wilcoxon rank sum test was used to calculate significance of differences between groups; and significance of the differences in proportions between groups was tested using the Fisher’s exact test.

3.3.4. TB and HIV comorbidities in the study population

New cases of Diabetes, TB and HIV were calculated for each year from 1st January 2011 to 31st December 2017. New Diabetes cases per year were counted as those with date of diabetes first ascertainment in that year, and this time range reflects a period during which most of the electronic data sources of the PHDC were in common use with relatively complete mortality data. New TB cases in each year were inferred from the PHDC episodes data as those where TB episode date was after date of diabetes first ascertainment. Likewise, new HIV cases were those where the date of HIV ascertainment was after diabetes ascertainment. The incidence of both TB and HIV per year in the study population over that 6-year period was calculated from these numbers.

Summary statistics describe the study population with a history of TB, comparing individuals with a TB episode before Diabetes was ascertained and those who developed active TB after their diabetes was ascertained (TB-Diabetes comorbidity). A person can have multiple cases of TB in their lifetime, and each time they are ascertained with TB it is recorded in the PHDC as a new TB episode with a start date and an end date. The ascertainment of TB episodes in relation to diabetes ascertainment was inferred using the episodes data to generate counts of the TB episodes of individuals before and after Diabetes ascertainment.

The HIV and TB status of the individual at the time of each HbA1c test was calculated and the 'time to HbA1c ascertainment relative to TB ascertainment' was inferred from the PHDC data by calculating the time difference in years between when each HbA1c test was done and when TB was ascertained in that individual. Negative time values were for HbA1c tests done before TB ascertainment, and positive time values were for HbA1c tests done after TB ascertainment.

The 'time to HbA1c ascertainment relative to HIV ascertainment' was inferred from the PHDC data by calculating the time difference in years between when each HbA1c test was done and when HIV status was ascertained in that individual. Negative time values were for HbA1c tests done before HIV ascertainment, and positive time values were for HbA1c tests done after someone is diagnosed as HIV positive.

3.3.5. Diabetes treatment

Counts of the different diabetes drugs of individuals who had ever started diabetes treatment were done and stratified according to diabetes duration of the study participants. The Chi-squared test measured statistical significance in the difference in the proportions of people who were in the different groups.

3.4. Results

3.4.1. The study population

There were 16 969 individuals with an inferred diabetes episode, and of these, 15 842 were identified as most likely having Type 2 Diabetes Mellitus (T2DM) according to the described inclusion/exclusion criteria. Of the individuals with T2DM, 13 771 had recorded HbA1c laboratory results and 13 528 had pharmacy records for diabetes medications in the PHDC routine data (Figure 3.1). Of the study population, 67% had an average of one HbA1c test annually for the years assessed, although timing/spacing of the tests was not consistent.

Summary statistics (Table 3.1) show the study population was 67% female, with a median age at diabetes ascertainment of 52 years (IQR: 44, 59) and a 58% (N = 8003) people had been ascertained with diabetes less than 5 years. Diabetes is a progressive disease and we saw the median HbA1c was higher if the period since diabetes ascertainment was longer, with those ascertained more than 10 years previously having significantly higher HbA1c (Table 3.1). Almost everyone (>99%) who had had diabetes for more than 5 years was on diabetes treatments compared to only 75.1% ($p<0.001$) in those were ascertained less than

5 years earlier (Table 3.1). In addition, 18% were ascertained as HIV-positive, and there was no significant difference in the proportion of HIV-positive individuals when considering how long they have had diabetes. There was, however, a significant difference in the proportions of people who had a history of TB, where those who have had diabetes for less than 5 years had the lowest proportion (11.4%) and those who have had diabetes for 10 years or more having the highest proportion of people (17.9%). The same trend was observed for those who had a TB-Diabetes comorbidity, where 91.8% of those with a history of TB who have had diabetes 10 years or more had an active TB episode after being ascertained with diabetes (Table 3.1).

Table 3.1. Characteristics of the whole study population, and stratified by duration of diabetes in years

	All *N=13771*	0- 5 Years *N=8003*	5-10 Years *N=5219*	≥ 10 Years *N=549*	p-value
Sex (Female)	9246 (67.2%)	5225 (65.4%)	3635 (69.7%)	386 (70.4%)	<0.001
Age at diabetes ascertainment (Years)	52.0 [44.0;59.0]	52.0 [44.0;61.0]	51.0 [44.0;58.0]	50.0 [42.0;57.0]	<0.001
Baseline HbA1c (%) [a]	8.5 [7.0;11.1]	7.9 [6.8;10.5]	9.5 [7.6;11.8]	9.9 [8.3;11.5]	<0.001
Last HbA1c (%)	9.5 [7.2;12.7]	8.6 [6.8;12.4]	10.3 [8.0;13.1]	10.8 [8.6;13.2]	<0.001
Patient outcome (Deceased)	631 (4.6%)	377 (4.7%)	237 (4.5%)	17 (3.1%)	0.213
Diabetes duration since ascertainment (Years)	4.1 [1.2;6.5]	1.6 [0.1;3.3]	6.6 [5.9;7.8]	10.7 [10.3;11.2]	0.000
Ever started diabetes treatment	11745 (85.3%)	6012 (75.1%)	5186 (99.4%)	547 (99.6%)	0.000
Ever had Tuberculosis	1839 (13.4%)	910 (11.4%)	831 (15.9%)	98 (17.9%)	<0.001
TB-Diabetes comorbidity [b]	1008 (55.9%)	372 (41.9%)	547 (67.0%)	89 (91.8%)	<0.001
HIV Positive	2508 (18.2%)	1478 (18.5%)	932 (17.9%)	98 (17.9%)	0.657

a. Baseline HbA1c was defined as the first recorded HbA1c value either at diabetes ascertainment or within the first year after diabetes ascertainment.
b. Proportions calculated from those who had ever had Tuberculosis

3.4.2. TB and HIV in people living with diabetes

Comparing HIV-positive and HIV-negative groups showed people living with HIV (PLWH) had diabetes ascertained at a significantly younger median age than the HIV-negative population (46 years c.f. 53 years respectively, p<0.001) (Table 3.2). In addition, PLWH had a significantly higher most recent HbA1c than the HIV-negative population (12.1% c.f. 9.1%, p<0.001). In line with other findings, the percentage of people who have ever had TB was significantly higher amongst PLWH (32% vs 9%), but the proportion of HIV-negative individuals who developed active TB after diabetes ascertainment was significantly higher (62% c.f. 48%, p<0.001) than for PLWH. There was a significantly higher proportion of people with a history of TB in those who were ascertained with HIV before diabetes (35.9% c.f. 26.5%, p<0.001) compared to those who were ascertained HIV after diabetes.; and there was a significantly higher proportion of people with a TB-Diabetes comorbidity (88.5% c.f. 30.1%, p<0.001) in those who were ascertained with HIV after diabetes compared to those who were ascertained HIV before diabetes (Supplementary Table 3.1). In addition, there was a significant difference in the percentage of people who were deceased when comparing those ascertained with HIV before or after diabetes ascertainment (4.5% c.f. 8.1%, p<0.001). This is unlikely to be only an effect of age, as the median ages at diabetes ascertainment in these two groups are 45.0 (IQR: 39.0, 52.0) c.f. 47.0 (IQR: 39.0, 53.0) years.

Table 3.2. Characteristics of the whole study population, and stratified by the HIV status of the participants

	ALL *N=13771*	HIV Negative *N=11263* (82%)	HIV Positive *N=2508* (18%)	p-value
Sex (Female)	9246 (67.2%)	7520 (66.9%)	1726 (68.9%)	0.054
Age at diabetes ascertainment (Years)	52.0 [44.0;59.0]	53.0 [46.0;61.0]	46.0 [39.0;52.0]	<0.001
Age categories				<0.001
18-39	2110 (15.3%)	1445 (12.8%)	665 (26.5%)	
40-49	3718 (27.0%)	2742 (24.3%)	976 (38.9%)	
50-59	4576 (33.2%)	3906 (34.7%)	670 (26.7%)	
60-69	2357 (17.1%)	2191 (19.5%)	166 (6.6%)	
70-79	805 (5.8%)	779 (6.9%)	26 (1.0%)	
>=80	205 (1.5%)	200 (1.8%)	5 (0.2%)	
Baseline HbA1c (%)	8.5 [7.0;11.1]	8.6 [7.0;11.1]	8.4 [6.9;10.9]	0.008
Baseline HbA1c < 7%	2820 (24.5%)	2268 (24.0%)	552 (26.4%)	0.023
Last HbA1c (%)	9.5 [7.2;12.7]	9.1 [7.1;12.0]	12.1 [7.9;15.0]	<0.001
Last HbA1c < 7%	2928 (21.3%)	2509 (22.3%)	419 (16.7%)	<0.001
Patient outcome (Deceased)	631 (4.6%)	488 (4.3%)	143 (5.7%)	0.004
Ever started diabetes treatment	11745 (85.3%)	9631 (85.5%)	2114 (84.3%)	0.126
Linkage to diabetes [a] treatment	10707 (91.2%)	8913 (92.5%)	1794 (84.9%)	<0.001

	ALL *N=13771*	HIV Negative *N=11263* (82%)	HIV Positive *N=2508* (18%)	p-value
Linkage to HbA1c testing [b]	9264 (67.3%)	7580 (67.3%)	1684 (67.2%)	0.909
Ever had Tuberculosis	1839 (13.4%)	1039 (9.2%)	800 (31.9%)	<0.001
Tuberculosis-Diabetes comorbidity	1008 (55.9%)	627 (62.0%)	381 (48.2%)	<0.001

a. Proportions of patients having a record of diabetes drugs being prescribed within one year of their last recorded diabetes-related hospital encounter (calculated from those who had ever started diabetes treatment)
b. Proportion of patients having a recorded HbA1c laboratory test result within one year of the last recorded diabetes related hospital encounter

The TB population (Supplementary Table 3.1) was 57% female with a median age at diabetes ascertainment of 49 years, and everyone in this cohort diagnosed with TB was linked to TB treatment. In addition, people with a history of TB had worse outcomes as we saw significantly more deceased people in this group when compared to those without a history of active TB disease (10% c.f. 3.8%, p<0.001). There was no significant difference in the gender distribution or age at diabetes ascertainment between those ascertained with TB before or after diabetes ascertainment (Table 3.3). The median baseline HbA1c of 10.1% (IQR: 7.6, 12.3) was significantly higher (p-value < 0.001) in those diagnosed with TB after diabetes when compared to those diagnosed with TB before diabetes at 8.2% (IQR: 6.8, 11.0). The results also suggest that developing active TB after a diabetes diagnosis may result in worse outcomes, as significantly more people in this group died (12%) when compared to those who had TB before being diagnosed with diabetes (7%). This is unlikely to be only an effect of age, as the median ages at diabetes ascertainment in these two groups are 49.0 (IQR: 41.2, 57.0) c.f. 48.0 (IQR: 41.0, 56.0) years.

Table 3.3. Characteristics of the study population with a history of Tuberculosis (TB) and stratified by the onset of the TB episode in relation to diabetes ascertainment

	ALL N=1802 (98%) [a]	TB episode before diabetes ascertainment N=794 (44%)	TB episode after diabetes ascertainment N=1008 (56%)	p-value
Sex (Female)	1023 (56.8%)	442 (55.7%)	581 (57.7%)	0.415
Age at diabetes ascertainment (Years)	49.0 [41.0;56.0]	49.0 [41.2;57.0]	48.0 [41.0;56.0]	0.13
Baseline HbA1c (%)	9.2 [7.1;11.8]	8.2 [6.8;11.0]	10.1 [7.6;12.4]	<0.001
Baseline Hba1c < 7%	327 (22.3%)	189 (28.9%)	138 (17.0%)	<0.001
Last HbA1c (%)	11.0 [7.6;14.1]	10.8 [7.2;14.4]	11.1 [8.0;13.9]	0.130

	ALL N=1802 (98%) [a]	TB episode before diabetes ascertainment N=794 (44%)	TB episode after diabetes ascertainment N=1008 (56%)	p-value
Last HbA1c < 7%	328 (18.2%)	173 (21.8%)	155 (15.4%)	<0.001
HIV Positive	791 (43.9%)	410 (51.6%)	381 (37.8%)	<0.001
Diabetes duration since ascertainment (Years)	5.1 [1.8;7.1]	3.3 [0.5;6.1]	6.1 [3.5;7.9]	<0.001
Patient outcome (Deceased)	178 (9.9%)	53 (6.7%)	125 (12.4%)	<0.001
Ever started TB treatment	1802 (100.0%)	794 (100.0%)	1008 (100.0%)	.
Ever started diabetes treatment	1579 (87.6%)	640 (80.6%)	939 (93.2%)	<0.001
Linkage to diabetes treatment [b]	1350 (85.5%)	536 (83.8%)	814 (86.7%)	0.120
Linkage to HbA1c testing [c]	1209 (67.1%)	532 (67.0%)	677 (67.2%)	0.983

a. 35 (2%) individuals who had ever had TB did not have enough data to classify when they had a TB episode relative to diabetes ascertainment
b. Proportions of patients having a record of diabetes drugs being prescribed within one year of their last recorded diabetes-related hospital encounter (calculated from those who had ever started diabetes treatment)
c. Proportion of patients having a recorded HbA1c laboratory test result within one year of the last recorded diabetes related hospital encounter

3.4.3. Annual incidence of TB and HIV

New cases of Diabetes, TB and HIV were calculated in each year from 1st January 2011 to 31st December 2017. There was a steady increase of newly ascertained diabetes cases over the six-year period excluding 2012 and 2013 (Figure 3.2). The data also show there were almost equal numbers of new TB and HIV cases in the study population, and these numbers steadily decreased over the six-year period except for 2012 in which there was a spike for both. The TB and HIV incidence in this diabetes population were calculated at 1.06% per year and 0.98% per year respectively, calculated over the six-year period.

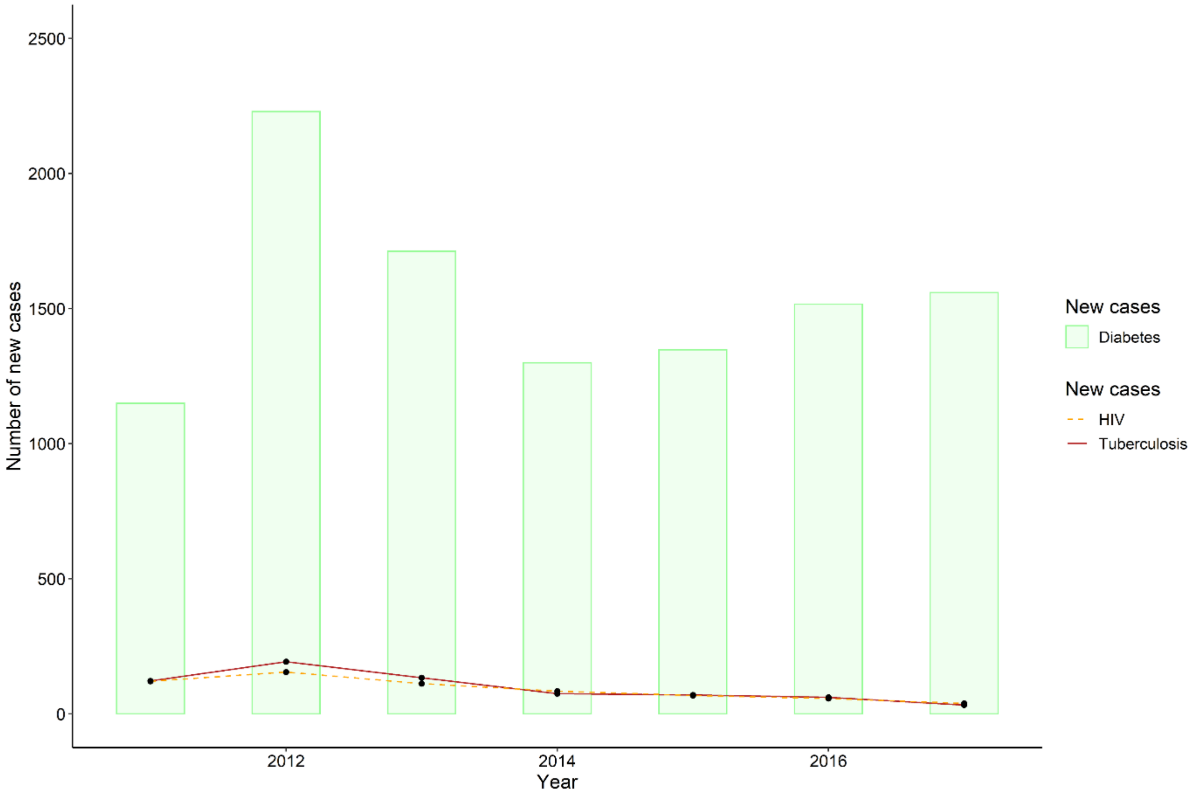


Figure 3.2. Bar graph showing new diabetes cases (bars) from January 2011 to December 2017 overlaid with line plots of new Tuberculosis (solid line) and HIV (dashed line) cases in these diabetes patients in the same time period.

3.4.4. Multiple episodes of TB

A person can have multiple cases of TB in their lifetime. Each time they are ascertained with TB it is recorded in the PHDC as a TB episode with start and end dates, and the ascertainment of TB episodes in relation to diabetes ascertainment was inferred using these data. There was a statistically significant difference (p-value < 0.001) in the distribution of TB episodes ascertained before and after diabetes ascertainment (Figure 3.3). The data show that after their first TB episode, significantly more people were getting subsequent TB episodes after diabetes ascertainment.

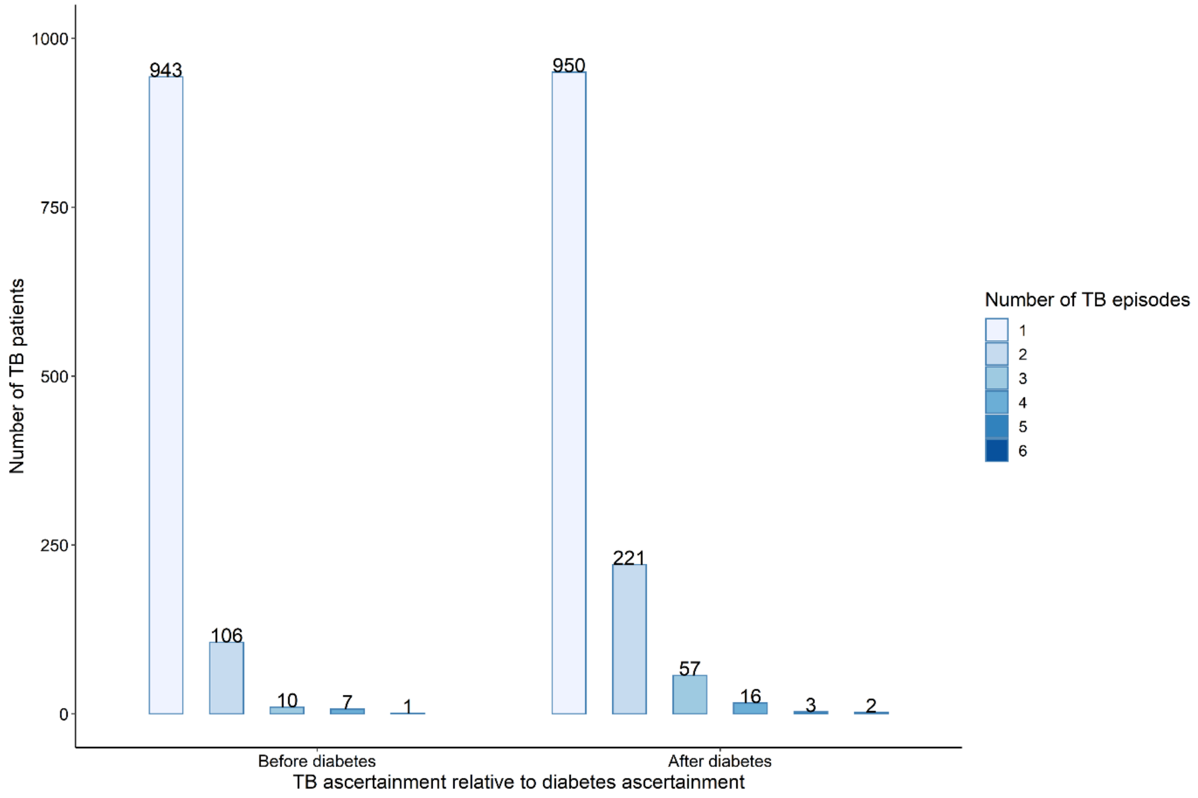


Figure 3.3. Distribution of repeat Tuberculosis (TB) episodes in the study population before and after diabetes ascertainment

3.4.5. HbA1c before and after TB ascertainment

The overall mean population HbA1c measured during both the 5 years before and 5 years after TB ascertainment is greater than 9% and is higher at the longer times since TB diagnosis, despite the majority of these patients receiving diabetes treatment (Figure 3.4A). Most of the HbA1c values of patients not on diabetes treatment are concentrated around an HbA1c of 6.5% which is the cut off HbA1c value for diagnosing diabetes, so it is reasonable to assume that these individuals are not yet receiving dispensed diabetes medications. Immediately after TB ascertainment, however, mean HbA1c is lower and there are more HbA1c values below 6.5% when compared to before TB ascertainment. In addition, after TB ascertainment, there are more recorded HbA1c values of patients not on diabetes treatment, and while most of these HbA1c values are concentrated around 6.5%, there are patients with HbA1c greater than 9% who are not on diabetes treatment. Both before and after TB ascertainment there is no distinct pattern for HbA1c values of patients who have had one or two TB episodes (Figure 3.4B), but for patients who have had three or four TB episodes the HbA1c values are mostly greater than 9%, and this is true for both before and after TB ascertainment. HbA1c values of participants who were deceased at study end were distributed randomly across the different HbA1c ranges both before and after TB ascertainment (Supplementary Figure 3.1).

3.4.6. HbA1c values with respect to HIV ascertainment

The overall mean HbA1c measured during the 5 years before HIV ascertainment is greater than 9% but is generally lower at later time points and generally lowest (less than 9%) immediately after HIV ascertainment (Figure 3.5A). After HIV ascertainment, however, the overall mean HbA1c is generally a bit higher at later time points averaging just above 9%. Before HIV ascertainment the HbA1c values of those who are not on diabetes treatment are concentrated around the 6.5 % diabetes diagnosis threshold, however after HIV ascertainment the HbA1c values of those not on diabetes treatment are distributed randomly across the different HbA1c values (Figure 3.5A). After HIV ascertainment, there were more recorded HbA1c values in individuals who have had TB, and the HbA1c values of those patients were distributed randomly across the different HbA1c ranges both before and after HIV ascertainment (Figure 3.5B). Similarly, HbA1c values of participants who were deceased at the study end were distributed randomly across the different HbA1c ranges both before and after HIV ascertainment (Supplementary Figure 3.2).

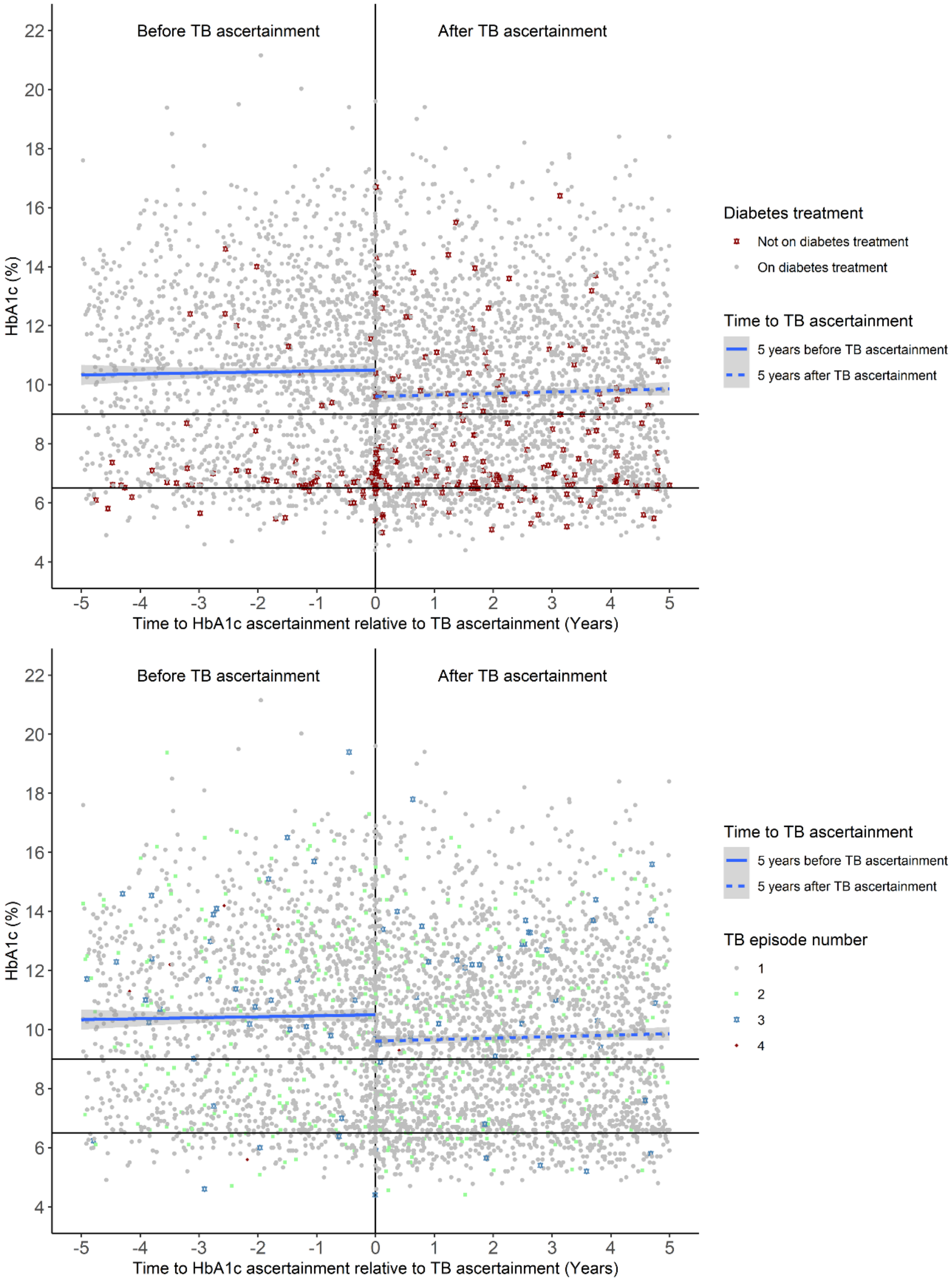


Figure 3.4. Effect of Tuberculosis ascertainment on HbA1c (%) over a 5-year period. A. HbA1c plotted by diabetes treatment i.e. on diabetes treatment (grey circle) or not on diabetes treatment (dark-red star). B. HbA1c plotted by TB episode i.e. 1 episode (grey circle), 2 episodes (pale-green square), 3 episodes (steel-blue star) or 4 episodes (dark-red diamond).

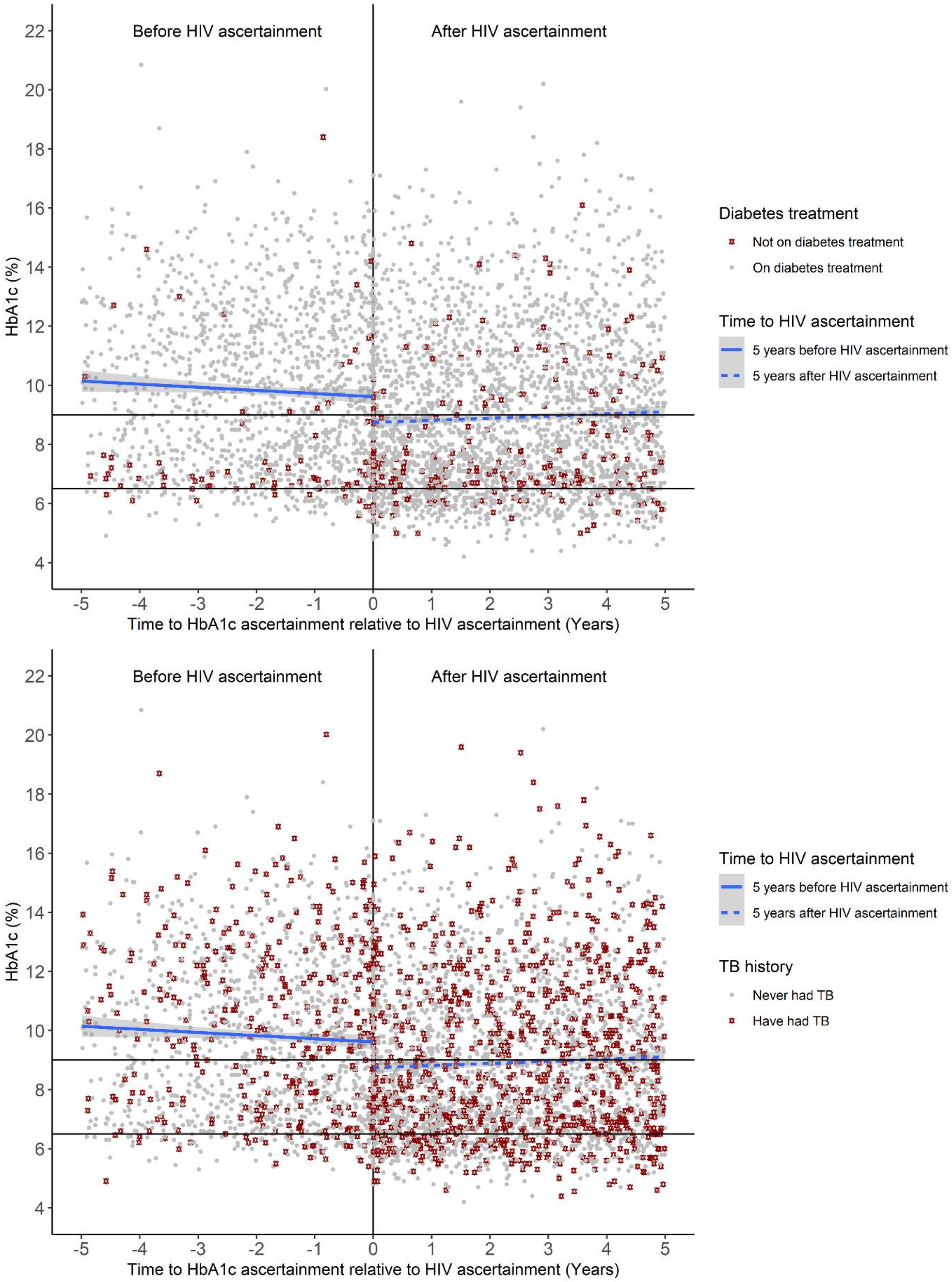


Figure 3.5. Effect of HIV ascertainment on HbA1c (%) over a 5-year period. A. HbA1c plotted by diabetes treatment i.e. on diabetes treatment (grey circle) or not on diabetes treatment (dark-red star). B. HbA1c plotted by TB history i.e. never had TB (grey circle) or have had TB (dark-red star).

3.4.7. Diabetes treatment

The study population was dispensed the three main diabetes drug classes available in the National Formulary for the Public Sector: Metformin (MTF), Sulphonylurea (SU) and insulin (Table 3.4). In line with current treatment practices, most of the population were on oral drugs and the most widely prescribed drug was metformin for 95% of the population, with 41% of treatment patients on insulin. In addition, the use of Insulin increased significantly with increasing duration of diabetes with 79.6% of the people who have had diabetes for more than 10 years on insulin (Table 3.4). This result is in line with the high median HbA1c of patients, but even though 85% of the study population was on treatment, the HbA1c was generally high and also seemed to be higher at later timepoints after diagnosis - suggesting that diabetes is failing to be controlled the longer patients have had the condition (Table 3.1).

Table 3.4. Pharmacy counts with last recorded HbA1c values for the whole population and stratified by duration of diabetes in years since ascertainment.

	ALL *N=13528*	**0 – 5 Years** *N=7748*	**5 – 10 Years** *N=5232*	**≥ 10 years** *N=548*	p-value
Metformin	12702 (95.2%)	7137 (94.4%)	5051 (96.6%)	514 (93.8%)	<0.001
Sulphonylurea	8309 (62.3%)	3846 (50.9%)	4046 (77.4%)	417 (76.1%)	<0.001
Insulin	5513 (41.3%)	2012 (26.6%)	3065 (58.6%)	436 (79.6%)	0.000
Metformin & Sulphonylurea	8684 (64.2%)	4093 (52.8%)	4150 (79.3%)	441 (80.5%)	<0.001
Metformin & Insulin	5078 (37.5%)	1745 (22.5%)	2927 (55.9%)	406 (74.1%)	0.000
Metformin, Insulin & Sulphonylurea	3778 (27.9%)	1039 (13.4%)	2416 (46.2%)	323 (58.9%)	0.000

Many PLWD also had TB and HIV comorbidities, and while all the TB and HIV patients in this study were recorded as having started treatment for each disease respectively, not all diabetes patients were on treatment for diabetes. For the PLWD with TB and HIV comorbidities, only 59.5% (1088) of those with a TB-DM comorbidity were recorded as being on treatment for both TB and diabetes simultaneously, while only 52.5% (1323) of those with an HIV-DM comorbidity were recorded as being on both HIV and diabetes treatment simultaneously. Only 40.6% (743) of patients with a triple TB-HIV-DM comorbidity were recorded as being on treatment for all three conditions simultaneously (Supplementary Table 3.3).

3.5. Discussion

The study population was drawn from individuals visiting health care facilities with over-representation of women, in line with other reports showing men are less likely to seek health care compared to women, and there is a general bias due to physically healthy women linking to health care through contraceptive and maternal health programmes whereas health men seldom visit health facilities (Abaerei et al., 2017). The distribution of people in the different age categories was similar for both men and women and the proportion of diabetes cases was highest at 33% in both men and women in the 50-59 age group (Supplementary Table 3.4). A significantly higher proportion of HIV positive people had diabetes ascertained at less than 50 years of age (65.4% vs 37.1%; p-value < 0.001) when compared to those who were HIV-negative at diabetes ascertainment (Table 3.2). Whilst there may be a causal relationship between HIV and diabetes, it is also possible that HIV positive people may have earlier ascertainment of diabetes because they are accessing care frequently and therefore getting screened and diagnosed earlier rather than diagnosis only happening once they develop severe symptoms. Median baseline HbA1c was similar for HIV positive 8.4% (IQR: 6.9, 10.9) and HIV negative groups 8.6% (IQR: 7.0, 11.1), suggesting that PLWH may be presenting with similar diabetes severity to HIV-negative patients at diagnosis. T2DM is a disease that is associated with ageing, but when comparing the HIV-positive and HIV-negative groups we saw a significantly higher proportion of people PLWH who were between 18-39 years (26.5 % c.f. 12.8% p-value < 0.001) being ascertained with T2DM (Table 3.2). This could be due to the interaction with HIV and Diabetes which increases the risk of diabetes and pre-diabetes in PLWH and especially those on highly active ART (HAART). There is also evidence that HIV significantly increases the risk of developing T2DM and that using highly active anti-retroviral therapy (HAART) induces hyperglycaemia (Araujo et al., 2014; Dave et al., 2011; Levitt et al., 2016), which is supported by our observations that, in a population with pre-existing diabetes, HIV co-infection appears in tandem with apparent glycaemic decline. We observed a median value of 8.4% (IQR: 6.9, 10.9) at baseline compared to 12.1% (IQR: 7.9, 15.0) at the last recorded HbA1c in this population, while in the HIV negative population there was only a slightly higher median of 9.1% (IQR: 7.1, 12.0) at the last recorded HbA1c compared to 8.6% (IQR: 7.0; 11.1) at baseline (Table 3.2). As all the HIV positive people in this study are on ART, the medications and the natural course of HIV infection might be contributing to the observed chronic hyperglycaemia. Other possibilities explanations include that HIV and diabetes care may not be well integrated in primary care clinics yet, and PLWH and DM may need to attend multiple clinics on multiple days leading to poor attendance.

The strong association between TB and HIV is well established and is reflected here with a TB burden in the HIV-positive population that is almost four times that in the HIV-negative population. Given the relationship between TB and HIV, a higher TB-Diabetes comorbidity in the HIV-positive group was expected, but we observed the opposite with significantly more HIV-negative people (62% c.f. 48.2%; p-value < 0.001) having a higher TB-Diabetes comorbidity (Table 1). This observation supports studies done in Nigeria (Lawson et al., 2017) and Tanzania (Faurholt-Jepsen et al., 2011) which showed that HIV negative people living with diabetes had an increased risk of developing pulmonary TB than HIV-positive people living with diabetes. It is estimated that up to 80% of the population in South Africa is infected with *Mycobacterium tuberculosis* however, not everyone who is infected progresses to TB disease (*WHO Global Tuberculosis Report 2019*, 2019). Studies have shown that in people with diabetes, the increased risk of TB disease is not necessarily from newly acquired infections, but rather by progression from latent to active TB (Koesoemadinata et al., 2017), however the biological mechanisms have not yet been elucidated. It is possible that in our study, the significantly higher proportions of HIV-negative people with TB could be driven by progression from latent to active TB disease caused by diabetes especially given that this population group is not put on TB preventive therapy, while it is part of clinical care in PLWH in South Africa (*WHO Global Tuberculosis Report 2019*, 2019). As the prevalence of diabetes continues to increase, it threatens to derail TB epidemic control efforts and there have been recent calls to assess the use of TB preventive therapy in people with diabetes (Harries, 2019; Jeon & Murray, 2008).

The relationship between T2DM and TB has been widely studied, but few studies have focused on the impact of active TB disease comorbidity on pre-existing diabetes. In this study we looked at the association between active TB disease and diabetes prognosis using HbA1c as an outcome. The target HbA1c for patients in care is 7% and as HbA1c levels increase so does the risk of diabetes complications (Sherwani et al., 2016). Results from our study show that in people with pre-existing diabetes, overall mean HbA1c is highest in the year before TB ascertainment and lowest in the year after (Figure 3.4). A possible explanation for this observation in our study population could be that the participants were linked to diabetes care following TB diagnosis resulting in an improvement in their diabetes control. It is also possible that having a TB diagnosis and subsequent in these individuals might result in better control of diabetes and improved HbA1c levels once they are not TB-positive. Because our data are routine health data and do not include any clinician notes, however, we cannot conclude this from these data alone. Even though the HbA1c was generally lower after TB ascertainment, it was greater that > 9% overall which is still classified as uncontrolled diabetes. Our results are not comparable to many other studies

(Aftab et al., 2017; Boillat-Blanco et al., 2016; Gupte et al., 2018; Tabarsi et al., 2014) because most of these studies were cross sectional or had a short follow up time and did not report HbA1c before TB ascertainment. In addition, the studies investigated the impact of TB on the diagnosis of new diabetes and not on pre-existing diabetes. Overall, we observed that having TB disease did not seem to influence the trajectory of glycaemic control in the long term, but PLWD who developed active TB had worse outcomes, as we saw significantly more deaths (12.4% vs 6.7% p-value < 0.001) in this group (Table 3.2). Previous studies have shown that TB patients diagnosed with diabetes have worse TB outcomes (Chiang et al., 2015; Wu et al., 2016) and the same seems to hold true for TB patients with pre-existing diabetes. Since survival was the only patient outcome measure used in this study beyond HbA1c values, we could not determine the impact of the observed chronic hyperglycaemia on risk of developing diabetes related vascular complications which were observed in other studies (Litwak et al., 2013).

HIV and diabetes are both chronic progressive illnesses which put a huge burden on the health care system (Williams et al., 2017), it is therefore important to understand how these two diseases affect each other in the South African context. While several studies have investigated the impact of HIV on glucose metabolism and the risk on developing pre-diabetes and diabetes (Araujo et al., 2014; Dave et al., 2011; Levitt et al., 2016), there is a paucity of studies investigating how HIV impacts the prognosis of pre-existing diabetes. In this study we aimed to investigate HbA1c levels in relation to new HIV infection in the context of pre-existing diabetes. Prior studies also show that HbA1c readings underestimate glycemia in HIV-infected individuals (Diop et al., 2006; P. S. Kim et al., 2009; S.-Y. Kim et al., 2014) and the results in our study might reflect these findings because we see a drop in mean HbA1c in the year following HIV ascertainment which only increases slightly over time. In addition, we also saw an overall trend in which HbA1c was lower before HIV ascertainment and this could be a possible indicator of undiagnosed HIV (Figure 3.5). It is also possible that the level of hyperglycaemia in PLWD who HIV are positive could be underestimated, suggesting that the utility of HbA1c in monitoring glycaemic control in HIV endemic settings like South Africa warrants further investigation.

T2DM can be managed using a combination of lifestyle changes and drug therapy and HbA1c levels are used as a proxy measure of long term diabetes control (“International Expert Committee Report on the Role of the A1C Assay in the Diagnosis of Diabetes,” 2009). An HbA1c < 7% is the target level for good glycaemic control (“Intensive Blood-Glucose Control with Sulphonylureas or Insulin Compared with Conventional Treatment and Risk of Complications in Patients with Type 2 Diabetes (UKPDS 33),” 1998), however studies have shown that worldwide, people living with diabetes are failing to reach this glycaemic target (Camara et al., 2015; Erasmus et al., 1999; Litwak et al., 2013; Musenge et al., 2015; Pinchevsky et al., 2015). This study had similar results with only 24.5% (n = 2820) of the study population showing good glycaemic control at baseline. While this is worrying, it reflects that more than two thirds of diabetes in SSA including in South Africa is undiagnosed (Mbanya et al., 2010) until patients present with symptoms of chronic hyperglycaemia. The aim of diabetes management is controlling hyperglycaemia to reduce the risk of progression to microvascular and macrovascular complications ("Intensive Blood-Glucose Control with Sulphonylureas or Insulin Compared with Conventional Treatment and Risk of Complications in Patients with Type 2 Diabetes (UKPDS 33)," 1998), but this study indicated that this population is failing to reach this target despite 85% being recorded as being on treatment. This is a worrying trend which is possibly due to a combination of diabetes disease progression with time, and a lack of compliance and adherence with the treatment and lifestyle changes [52,54–56]. Further analysis is needed to establish adherence and compliance in the study population, as this cannot be determined from the retrospective data alone.

HIV and diabetes are both chronic diseases whose long-term management includes drug therapy, however, only 52.8% of the study population with an HIV-diabetes comorbidity were on diabetes treatment. It is possible that there are both patient, provider and systems issues causing delay in initiation of therapy. Some patients might also get their diabetes care in the private health sector at different times during their care, and private health data were not included in this study, but it is unlikely that they would access public health facilities for one illness but not the other. These data suggest that a coordinated response is needed to address the gaps and provide an holistic, integrated care for people living with diabetes, especially in the context of the high burden of infectious diseases in Africa. Such an integrated approach would include education of PLWD, availability of health professionals with required skills, and sociodemographic considerations (Gennaro et al., 2019). It will be important to better understand why almost 50% of patients with HIV-diabetes comorbidity are not on diabetes treatment despite the high median HbA1c suggesting a need for treatment intervention, and prospective studies can explore factors that determine treatment timelines especially with associated HIV diagnosis.

3.6. Potential limitations of the study

There is a two-tier health system in South Africa where some individuals receive private health care, some receive only government health care, and there are also many individuals who access both types of service and transition back and forth depending on their employment and health insurance status (Chopra et al., 2009; Winchester & King, 2018). We therefore expect that an exhaustive health record for each individual may not be available through the PHDC. Some patients did not have recorded HbA1c results and pharmacy records, and this may could be due to private health service utilisation, as well as the staggered roll-out of electronic health data platforms in the Province which means that data completeness may fluctuate according to the facility attended and year of service provision. Also, South Africa has a federated health service whereby provinces manage healthcare services (Chopra et al., 2009; Winchester & King, 2018), and coupled with a highly migratory working population, it is possible that records are missing when individuals move to other provinces in South Africa for periods of time.

3.7. Conclusion

To our knowledge this is the first study in South Africa to use longitudinal routine health data to study the relationship between active TB disease and HIV infection in the context of pre-existing diabetes using National Glycohemoglobin Standardization Program (NGSP) HbA1c as an outcome measure. In addition, we were able to establish temporal order of disease ascertainment. The study had a large sample size and long-term retrospective data, reducing selection bias arising from including people actively seeking care. In addition, these routine health data reflect a more accurate picture of diabetes in the general population than would actively managed clinical studies involving diabetes patients. The epidemiologic findings in this exploratory study demonstrate that routine health data are a valuable resource for understanding disease epidemiology and highlighted the need for further research into diabetes outcomes in a high TB and HIV burden setting.

3.8. References

Abaerei, A. A., Ncayiyana, J., & Levin, J. (2017). Health-care utilization and associated factors in Gauteng province, South Africa. *Global Health Action*, *10*(1), 1305765. https://doi.org/10.1080/16549716.2017.1305765

Africa, S. S. (n.d.). *Publication | Statistics South Africa*. Retrieved January 20, 2021, from http://www.statssa.gov.za/?page_id=1854

Aftab, H., Christensen, D. L., Ambreen, A., Jamil, M., Garred, P., Petersen, J. H., Nielsen, S. D., & Bygbjerg, I. C. (2017). Tuberculosis-Related Diabetes: Is It Reversible after Complete Treatment? *The American Journal of Tropical Medicine and Hygiene*, *97*(4), 1099–1102. https://doi.org/10.4269/ajtmh.16-0816

Araujo, S., Bañón, S., Machuca, I., Moreno, A., Pérez-Elías, M. J., & Casado, J. L. (2014). Prevalence of insulin resistance and risk of diabetes mellitus in HIV-infected patients receiving current antiretroviral drugs. *European Journal of Endocrinology*, *171*(5), 545–554. https://doi.org/10.1530/EJE-14-0337

Bertram, M. Y., Jaswal, A. V. S., Wyk, V. P. V., Levitt, N. S., & Hofman, K. J. (2013). The non-fatal disease burden caused by type 2 diabetes in South Africa, 2009. *Global Health Action*, *6*(1), 19244. https://doi.org/10.3402/gha.v6i0.19244

Boillat-Blanco, N., Ramaiya, K. L., Mganga, M., Minja, L. T., Bovet, P., Schindler, C., Von Eckardstein, A., Gagneux, S., Daubenberger, C., Reither, K., & Probst-Hensch, N. (2016). Transient Hyperglycemia in Patients With Tuberculosis in Tanzania: Implications for Diabetes Screening Algorithms. *The Journal of Infectious Diseases*, *213*(7), 1163–1172. https://doi.org/10.1093/infdis/jiv568

Boulle, A., Heekes, A., Tiffin, N., Smith, M., Mutemaringa, T., Zinyakatira, N., Phelanyane, F., Pienaar, C., Buddiga, K., Coetzee, E., Rooyen, R. van, Dyers, R., Fredericks, N., Loff, A., Shand, L., Moodley, M., Vega, I. de, & Vallabhjee, K. (2019). Data Centre Profile: The Provincial Health Data Centre of the Western Cape Province, South Africa. *International Journal of Population Data Science*, *4*(2), Article 2. https://doi.org/10.23889/ijpds.v4i2.1143

Camara, A., Baldé, N. M., Sobngwi-Tambekou, J., André, P., Kengne, A. P., Diallo, M. M., Tchatchoua, A. P. K., Kaké, A., Ngamani, S., Balkau, B., Bonnet, F., & Sobngwi, E. (2015). Poor glycemic control in type 2 diabetes in the South of the Sahara: The issue of limited access to an HbA1c test. *Diabetes Research and Clinical Practice*, *108*(1), 187–192. https://doi.org/10.1016/j.diabres.2014.08.025

Cheng, J., Zhang, H., Zhao, Y. L., Wang, L. X., & Chen, M. T. (2017). Mutual Impact of Diabetes Mellitus and Tuberculosis in China. *Biomedical and Environmental Sciences: BES*, *30*(5), 384–389. https://doi.org/10.3967/bes2017.051

Chiang, C. Y., Bai, K. J., Lin, H. H., Chien, S. T., Lee, J. J., Enarson, D. A., Lee, T.-I., & Yu, M.-C. (2015). The Influence of Diabetes, Glycemic Control, and Diabetes-Related

Comorbidities on Pulmonary Tuberculosis. *PLOS ONE*, *10*(3), e0121698. https://doi.org/10.1371/journal.pone.0121698

Chopra, M., Daviaud, E., Pattinson, R., Fonn, S., & Lawn, J. E. (2009). Saving the lives of South Africa's mothers, babies, and children: Can the health system deliver? *The Lancet*, *374*(9692), 835–846. https://doi.org/10.1016/S0140-6736(09)61123-5

Dave, J. A., Lambert, E. V., Badri, M., West, S., Maartens, G., & Levitt, N. S. (2011). Effect of Nonnucleoside Reverse Transcriptase Inhibitor–Based Antiretroviral Therapy on Dysglycemia and Insulin Sensitivity in South African HIV-Infected Patients. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, *57*(4), 284–289. https://doi.org/10.1097/QAI.0b013e318221863f

Diop, M.-E., Bastard, J.-P., Meunier, N., Thévenet, S., Maachi, M., Capeau, J., Pialoux, G., & Vigouroux, C. (2006). Inappropriately low glycated hemoglobin values and hemolysis in HIV-infected patients. *AIDS Research and Human Retroviruses*, *22*(12), 1242–1247. https://doi.org/10.1089/aid.2006.22.1242

Eckhardt, B. J., Holzman, R. S., Kwan, C. K., Baghdadi, J., & Aberg, J. A. (2012). Glycated Hemoglobin A(1c) as screening for diabetes mellitus in HIV-infected individuals. *AIDS Patient Care and STDs*, *26*(4), 197–201. https://doi.org/10.1089/apc.2011.0379

Erasmus, R. T., Blanco, E. B., Okesina, A. B., Gqweta, Z., & Matsha, T. (1999). Assessment of glycaemic control in stable type 2 black South African diabetics attending a peri-urban clinic. *Postgraduate Medical Journal*, *75*(888), 603–606. https://doi.org/10.1136/pgmj.75.888.603

Faurholt-Jepsen, D., Range, N., PrayGod, G., Jeremiah, K., Faurholt-Jepsen, M., Aabye, M. G., Changalucha, J., Christensen, D. L., Pipper, C. B., Krarup, H., Witte, D. R., Andersen, A. B., & Friis, H. (2011). Diabetes Is a Risk Factor for Pulmonary Tuberculosis: A Case-Control Study from Mwanza, Tanzania. *PLOS ONE*, *6*(8), e24215. https://doi.org/10.1371/journal.pone.0024215

García-Pérez, L.-E., Álvarez, M., Dilla, T., Gil-Guillén, V., & Orozco-Beltrán, D. (2013). Adherence to Therapies in Patients with Type 2 Diabetes. *Diabetes Therapy*, *4*(2), 175–194. https://doi.org/10.1007/s13300-013-0034-y

Gennaro, F. D., Marotta, C., Antunes, M., & Pizzol, D. (2019). Diabetes in active tuberculosis in low-income countries: To test or to take care? *The Lancet Global Health*, *7*(6), e707. https://doi.org/10.1016/S2214-109X(19)30173-1

Global tuberculosis report 2020. (n.d.). Retrieved March 27, 2021, from https://www.who.int/publications-detail-redirect/9789240013131

Gouda, H. N., Charlson, F., Sorsdahl, K., Ahmadzada, S., Ferrari, A. J., Erskine, H., Leung, J., Santamauro, D., Lund, C., Aminde, L. N., Mayosi, B. M., Kengne, A. P., Harris, M., Achoki, T., Wiysonge, C. S., Stein, D. J., & Whiteford, H. (2019). Burden of non-communicable diseases in sub-Saharan Africa, 1990-2017: Results from the Global Burden of Disease Study 2017. *The Lancet. Global Health*, *7*(10), e1375–e1387. https://doi.org/10.1016/S2214-109X(19)30374-2

Gupte, A. N., Mave, V., Meshram, S., Lokhande, R., Kadam, D., Dharmshale, S., Bharadwaj, R., Kagal, A., Pradhan, N., Deshmukh, S., Atre, S., Sahasrabudhe, T., Barthwal, M., Meshram, S., Kakrani, A., Kulkarni, V., Raskar, S., Suryavanshi, N., Shivakoti, R., … Golub, J. E. (2018). Trends in HbA1c levels and implications for diabetes screening in tuberculosis cases undergoing treatment in India. *The International Journal of Tuberculosis and Lung Disease : The Official Journal of the International Union against Tuberculosis and Lung Disease*, *22*(7), 800–806. https://doi.org/10.5588/ijtld.18.0026

Hall, V., Thomsen, R. W., Henriksen, O., & Lohse, N. (2011). Diabetes in Sub Saharan Africa 1999-2011: Epidemiology and public health implications. a systematic review. *BMC Public Health*, *11*(1), 564. https://doi.org/10.1186/1471-2458-11-564

Harries, A. D. (2019). Having diabetes and being underweight in Asia: A potent risk factor for tuberculosis. *The International Journal of Tuberculosis and Lung Disease*, *23*(12), 1237–1238. https://doi.org/10.5588/ijtld.19.0599

Harries, A. D., Kumar, A. M. V., Satyanarayana, S., Lin, Y., Zachariah, R., Lönnroth, K., & Kapur, A. (2015). Diabetes mellitus and tuberculosis: Programmatic management issues. *The International Journal of Tuberculosis and Lung Disease*, *19*(8), 879–886. https://doi.org/10.5588/ijtld.15.0069

Harries, A. D., Lin, Y., Satyanarayana, S., Lönnroth, K., Li, L., Wilson, N., Chauhan, L. S., Zachariah, R., Baker, M. A., Jeon, C. Y., Murray, M. B., Maher, D., Bygbjerg, I. C., Enarson, D. A., Billo, N. E., & Kapur, A. (2011). The looming epidemic of diabetes-associated tuberculosis: Learning lessons from HIV-associated tuberculosis. *The International Journal of Tuberculosis and Lung Disease*, *15*(11), 1436–1445. https://doi.org/10.5588/ijtld.11.0503

IDF Diabetes Atlas 9th edition 2019. (n.d.). Retrieved February 20, 2021, from https://www.diabetesatlas.org/en/

Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). (1998). *The Lancet*, *352*(9131), 837–853. https://doi.org/10.1016/S0140-6736(98)07019-6

International Expert Committee Report on the Role of the A1C Assay in the Diagnosis of Diabetes. (2009). *Diabetes Care*, *32*(7), 1327. https://doi.org/10.2337/dc09-9033

Jeon, C. Y., & Murray, M. B. (2008). Diabetes Mellitus Increases the Risk of Active Tuberculosis: A Systematic Review of 13 Observational Studies. *PLoS Medicine*, *5*(7). https://doi.org/10.1371/journal.pmed.0050152

Khunti, N., Khunti, N., & Khunti, K. (2019). Adherence to type 2 diabetes management. *British Journal of Diabetes*, *19*(2), Article 2. https://doi.org/10.15277/bjd.2019.223

Kim, P. S., Woods, C., Georgoff, P., Crum, D., Rosenberg, A., Smith, M., & Hadigan, C. (2009). A1C Underestimates Glycemia in HIV Infection. *Diabetes Care*, *32*(9), 1591–1593. https://doi.org/10.2337/dc09-0177

Kim, S.-Y., Friedmann, P., Seth, A., & Fleckman, A. M. (2014). Monitoring HIV-infected Patients with Diabetes: Hemoglobin A1c, Fructosamine, or Glucose? *Clinical*

Medicine Insights. Endocrinology and Diabetes, *7*, 41–45. https://doi.org/10.4137/CMED.S19202

Koesoemadinata, R. C., McAllister, S. M., Soetedjo, N. N. M., Febni Ratnaningsih, D., Ruslami, R., Kerry, S., Verrall, A. J., Apriani, L., van Crevel, R., Alisjahbana, B., & Hill, P. C. (2017). Latent TB infection and pulmonary TB disease among patients with diabetes mellitus in Bandung, Indonesia. *Transactions of The Royal Society of Tropical Medicine and Hygiene*, *111*(2), 81–89. https://doi.org/10.1093/trstmh/trx015

Lawson, L., Muc, M., Oladimeji, O., Iweha, C., Opoola, B., Abdurhaman, S. T., Bimba, J. S., & Cuevas, L. E. (2017). Tuberculosis and diabetes in Nigerian patients with and without HIV. *International Journal of Infectious Diseases*, *61*, 121–125. https://doi.org/10.1016/j.ijid.2017.06.014

Leung, C. C., Lam, T. H., Chan, W. M., Yew, W. W., Ho, K. S., Leung, G. M., Law, W. S., Tam, C. M., Chan, C. K., & Chang, K. C. (2008). Diabetic control and risk of tuberculosis: A cohort study. *American Journal of Epidemiology*, *167*(12), 1486–1494. https://doi.org/10.1093/aje/kwn075

Levitt, N. S. (2008). Diabetes in Africa: Epidemiology, management and healthcare challenges. *Heart*, *94*(11), 1376–1382. https://doi.org/10.1136/hrt.2008.147306

Levitt, N. S., Peer, N., Steyn, K., Lombard, C., Maartens, G., Lambert, E. V., & Dave, J. A. (2016). Increased risk of dysglycaemia in South Africans with HIV; especially those on protease inhibitors. *Diabetes Research and Clinical Practice*, *119*, 41–47. https://doi.org/10.1016/j.diabres.2016.03.012

Litwak, L., Goh, S.-Y., Hussein, Z., Malek, R., Prusty, V., & Khamseh, M. E. (2013). Prevalence of diabetes complications in people with type 2 diabetes mellitus and its association with baseline characteristics in the multinational A1chieve study. *Diabetology & Metabolic Syndrome*, *5*(1), 57. https://doi.org/10.1186/1758-5996-5-57

Mbanya, J. C. N., Motala, A. A., Sobngwi, E., Assah, F. K., & Enoru, S. T. (2010). Diabetes in sub-Saharan Africa. *The Lancet*, *375*(9733), 2254–2266. https://doi.org/10.1016/S0140-6736(10)60550-8

Munseri, P. J., Kimambo, H., & Pallangyo, K. (2019). Diabetes mellitus among patients attending TB clinics in Dar es Salaam: A descriptive cross-sectional study. *BMC Infectious Diseases*, *19*. https://doi.org/10.1186/s12879-019-4539-5

Musenge, E. M., Michelo, C., Mudenda, B., & Manankov, A. (2015, December 21). *Glycaemic Control and Associated Self-Management Behaviours in Diabetic Outpatients: A Hospital Based Observation Study in Lusaka, Zambia* [Research Article]. Journal of Diabetes Research; Hindawi. https://doi.org/10.1155/2016/7934654

Nansseu, J. R., Bigna, J. J., Kaze, A. D., & Noubiap, J. J. (2018). Incidence and Risk Factors for Prediabetes and Diabetes Mellitus Among HIV-infected Adults on Antiretroviral Therapy: A Systematic Review and Meta-analysis. *Epidemiology*, *29*(3), 431–441. https://doi.org/10.1097/EDE.0000000000000815

Noncommunicable Diseases. (n.d.). WHO | Regional Office for Africa. Retrieved February 25, 2021, from https://www.afro.who.int/health-topics/noncommunicable-diseases

Oni, T., Berkowitz, N., Kubjane, M., Goliath, R., Levitt, N. S., & Wilkinson, R. J. (2017). Trilateral overlap of tuberculosis, diabetes and HIV-1 in a high-burden African setting: Implications for TB control. *European Respiratory Journal*, *50*(1). https://doi.org/10.1183/13993003.00004-2017

Pinchevsky, Y., Shukla, V., Butkow, N., Raal, F. J., & Chirwa, T. (2015). The achievement of glycaemic, blood pressure and LDL cholesterol targets in patients with type 2 diabetes attending a South African tertiary hospital outpatient clinic. *Journal of Endocrinology, Metabolism and Diabetes of South Africa*, *20*(2), 81–86. https://doi.org/10.1080/16089677.2015.1056468

Pizzol, D., Di Gennaro, F., Chhaganlal, K. D., Fabrizio, C., Monno, L., Putoto, G., & Saracino, A. (2017). Prevalence of diabetes mellitus in newly diagnosed pulmonary tuberculosis in Beira, Mozambique. *African Health Sciences*, *17*(3), 773–779. https://doi.org/10.4314/ahs.v17i3.20

Pizzol, D., Veronese, N., Marotta, C., Di Gennaro, F., Moiane, J., Chhaganlal, K., Monno, L., Putoto, G., Mazzucco, W., & Saracino, A. (2018). Predictors of therapy failure in newly diagnosed pulmonary tuberculosis cases in Beira, Mozambique. *BMC Research Notes*, *11*(1), 99. https://doi.org/10.1186/s13104-018-3209-9

Polonsky, W. H., & Henry, R. R. (2016). Poor medication adherence in type 2 diabetes: Recognizing the scope of the problem and its key contributors. *Patient Preference and Adherence*, *10*, 1299–1307. https://doi.org/10.2147/PPA.S106821

Rasmussen, L. D., May, M. T., Kronborg, G., Larsen, C. S., Pedersen, C., Gerstoft, J., & Obel, N. (2015). Time trends for risk of severe age-related diseases in individuals with and without HIV infection in Denmark: A nationwide population-based cohort study. *The Lancet. HIV*, *2*(7), e288-298. https://doi.org/10.1016/S2352-3018(15)00077-6

Reid, M., McFadden, N., & Tsima, B. (2013). Clinical challenges in the co-management of diabetes mellitus and tuberculosis in southern Africa. *Journal of Endocrinology, Metabolism and Diabetes of South Africa*, *18*(3), 135–140. https://doi.org/10.1080/22201009.2013.10844551

Restrepo, B. I. (2007). Convergence of the tuberculosis and diabetes epidemics: Renewal of old acquaintances. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, *45*(4), 436–438. https://doi.org/10.1086/519939

Schouten, J., Wit, F. W., Stolte, I. G., Kootstra, N. A., van der Valk, M., Geerlings, S. E., Prins, M., Reiss, P., & AGEhIV Cohort Study Group. (2014). Cross-sectional comparison of the prevalence of age-associated comorbidities and their risk factors between HIV-infected and uninfected individuals: The AGEhIV cohort study. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, *59*(12), 1787–1797. https://doi.org/10.1093/cid/ciu701

Sherwani, S. I., Khan, H. A., Ekhzaimy, A., Masood, A., & Sakharkar, M. K. (2016). Significance of HbA1c Test in Diagnosis and Prognosis of Diabetic Patients. *Biomarker Insights*, *11*, 95–104. https://doi.org/10.4137/BMI.S38440

Stokes, A., Berry, K. M., Mchiza, Z., Parker, W., Labadarios, D., Chola, L., Hongoro, C., Zuma, K., Brennan, A. T., Rockers, P. C., & Rosen, S. (2017). Prevalence and unmet need for diabetes care across the care continuum in a national sample of South African adults: Evidence from the SANHANES-1, 2011-2012. *PLOS ONE*, *12*(10), e0184264. https://doi.org/10.1371/journal.pone.0184264

Tabarsi, P., Baghaei, P., Marjani, M., Vollmer, W. M., Masjedi, M.-R., & Harries, A. D. (2014). Changes in glycosylated haemoglobin and treatment outcomes in patients with tuberculosis in Iran: A cohort study. *Journal of Diabetes & Metabolic Disorders*, *13*(1), 123. https://doi.org/10.1186/s40200-014-0123-0

Use of glycated haemoglobin (HbA1c) in the diagnosis of diabetes mellitus. (2011). *Diabetes Research and Clinical Practice*, *93*(3), 299–309. https://doi.org/10.1016/j.diabres.2011.03.012

WHO Global Tuberculosis Report 2019. (2019, October 18). USAID TB South Africa Project. https://tbsouthafrica.org.za/resources/who-global-tuberculosis-report-2019

Williams, B. G., Gupta, S., Wollmers, M., & Granich, R. (2017). Progress and prospects for the control of HIV and tuberculosis in South Africa: A dynamical modelling study. *The Lancet Public Health*, *2*(5), e223–e230. https://doi.org/10.1016/S2468-2667(17)30066-X

Winchester, M. S., & King, B. (2018). Decentralization, healthcare access, and inequality in Mpumalanga, South Africa. *Health & Place*, *51*, 200–207. https://doi.org/10.1016/j.healthplace.2018.02.009

Workneh, M. H., Bjune, G. A., & Yimer, S. A. (2017). Prevalence and associated factors of tuberculosis and diabetes mellitus comorbidity: A systematic review. *PloS One*, *12*(4), e0175925. https://doi.org/10.1371/journal.pone.0175925

Wu, Z., Guo, J., Huang, Y., Cai, E., Zhang, X., Pan, Q., Yuan, Z., & Shen, X. (2016). Diabetes mellitus in patients with pulmonary tuberculosis in an aging population in Shanghai, China: Prevalence, clinical characteristics and outcomes. *Journal of Diabetes and Its Complications*, *30*(2), 237–241. https://doi.org/10.1016/j.jdiacomp.2015.11.014

4. Chapter 4 : Risk factors for COVID-19 hospitalisation and death in people living with diabetes: A virtual cohort study from the Western Cape Province, South Africa

Dave JA, Tamuhla T, Tiffin N, Levitt NS, Ross IL, Toet W, et al. Risk factors for COVID-19 hospitalisation and death in people living with diabetes: A virtual cohort study from the Western Cape Province, South Africa. Diabetes Res Clin Pract. 2021;177: 108925. doi:10.1016/j.diabres.2021.108925

Relevance of paper in thesis

The association between diabetes and infectious disease is well established and in the previous chapter we demonstrated the link between diabetes and the infectious diseases TB, and HIV. Similarly at the start of the COVID-19 pandemic, several large studies reported on the association of increased severity of COVID-19 with diabetes, but most of the studies were from countries in the global north. At the time of doing the analysis, there were no large studies with a focus on diabetes from Africa that had been reported. In this analysis we addressed this knowledge gap and highlighted the utility of routine health data in understanding the epidemiology of an emerging infectious disease. We were able to scale up the data analysis pipelines that had been used in the Khayelitsha cohort and applied them to a large cohort in the Western Cape Province. This analysis also addressed objective 2 where we further explored statistical methods for modelling longitudinal data so that it can be used as a GWAS phenotype.

4.1. Abstract

Background

COVID-19 outcomes and risk factors, including comorbidities and medication regimens, in people living with diabetes (PLWD) are poorly defined for low- and middle-income countries.

Methods

The Provincial Health Data Centre (Western Cape, South Africa) is a health information exchange collating patient-level routine health data for approximately 4 million public sector health care seekers. Data from COVID-19 patients diagnosed between March and July 2020, including PLWD, were analysed to describe risk factors, including dispensed diabetes medications and comorbidities, and their association with COVID-19 outcomes in this population.

Findings

There were 64,476 COVID-19 patients diagnosed. Of 9305 PLWD, 44.9% were hospitalised, 4.0% admitted to ICU, 0.6% received ventilation and 15.4% died. In contrast, proportions of COVID-19 patients without diabetes were: 12.2% hospitalised, 1.0% admitted, 0.1% ventilated and 4.6% died. PLWD were significantly more likely to be admitted (OR:3.73, 95 %CI: 3.53, 3.94) and to die (OR:3.01, 95 %CI: 2.76,3.28). Significant hospitalised risk factors included HIV infection, chronic kidney disease, current TB, male sex and increasing age. Significant risk factors for mortality were CKD, male sex, HIV infection, previous TB and increasing age. Pre-infection use of insulin was associated with a significant increased risk for hospitalisation (OR:1.39, 95 %CI:1.24,1.57) and mortality (OR:1.49, 95 % CI:1.27; 1.74) and metformin was associated with a reduced risk for hospitalisation (OR:0.2, 95 %CI: 0.55, 0.71) and mortality (OR 0.77, 95 %CI:0.64; 0.92).

Interpretation

Using routine health data from this large virtual cohort, we have described the association of infectious and noncommunicable comorbidities as well as pre-infection diabetes medications with COVID-19 outcomes in PLWD in the Western Cape, South Africa.

4.2. Introduction

Infection with SARS-CoV-2 has caused a global pandemic that has not spared any country. At the time of writing, 107 million people have been infected and 2.4 million people have died from Coronavirus Disease in 2019 (COVID-19) (*COVID Live - Coronavirus Statistics - Worldometer*, n.d.). It has now been well established that people living with diabetes (PLWD) are at increased risk of more severe infection with SARS-CoV-2 ("COVID-19," 2020). Observational data of COVID-19 from countries which were at the forefront of this pandemic have reported greater morbidity and mortality in PLWD than those without diabetes, when studied in well-defined populations from the UK (Barron et al., 2020; Holman et al., 2020; Mancia et al., 2020; McGurnaghan et al., 2021; Williamson et al., 2020). Publications from low-middle income countries of Mexico, Brazil and South Africa have similarly confirmed this risk (Bello-Chavolla et al., 2020; Martins-Filho et al., 2021; Western Cape Department of Health in collaboration with the National Institute for Communicable Diseases, 2021).

Determinants of COVID-19 mortality risk among people with diabetes have been explored in a few studies from the UK, Ireland and France (Barron et al., 2020; Cariou et al., 2020; Holman et al., 2020; McGurnaghan et al., 2021). Age, sex, duration of diabetes, body mass index (BMI), black and South Asian ethnicity, lower socio-economic status, poorer glycaemic control, and pre-existing cardiovascular disease were reported to increase risk.

In this study, we used linked-routine health data at the end of the first wave of COVID-19, collated from a variety of electronic platforms for adults attending public sector health facilities in the Western Cape Province, South Africa, to identify whether diabetes is associated with greater morbidity and mortality from COVID-19. By using routinely collected data, we aimed to determine whether there were any predictors for more severe COVID-19, among these patients.

4.3. Methods

4.3.1. Study design

In this cohort study, we used data from the first wave of the pandemic in the Western Cape Province, from 04 March 2020 (when the first case was identified) to 15 July 2020, when infection rates had dropped.

4.3.2. Selection of study population

The Provincial Health Data Centre (PHDC) is a health information exchange, housed by the Western Cape Department of Health that collates and links routine health data from a variety of electronic platforms used across the Western Cape Province (Boulle et al., 2019). These include demographic data from facilities, dispensing data for medications, laboratory data, and data from a variety of disease-specific and service delivery data systems. The data are updated daily and linked to a deduplicated patient master index (PMI), which represents approximately 5.25 million that rely solely on the public sector for health care.

The study population was identified from the Western Cape Population, as represented in the PHDC. Inclusion criteria were: (1) having attended at least one Government Health Facility in the Western Cape, South Africa, in the period 1 January 2010 to 31 December 2019, used as a proxy for patients accessing public sector health care and (2) a laboratory confirmation of SARS-CoV-2 infection up and until the 15 July 2020. A COVID-19 diagnosis was inferred from PHDC records, using evidence of a positive SARS-CoV-2 polymerase chain reaction (PCR) laboratory result. Records without patient COVID-19 outcome (deceased or recovered) by October 2020, were excluded.

Data descriptors: All retrospective, routine health data accessed in PHDC for public health sector patients with a COVID-19 outcome, were analysed using R version 3.6.1 (2019-07-05). Descriptive statistical methods assessed population demographics as of 31 July 2020. ‘Age’ was the age at COVID-19 diagnosis in years, ‘sex’ was the gender recorded in PHDC records (male, female), ‘pregnant’ indicated pregnancy status at COVID-19 diagnosis. ‘Hospital_admission’ referred to hospital admission contemporaneous with COVID-19, ‘admitted_to_ICU’ represents admission to an intensive care unit (ICU) due to COVID-19 and ‘ventilated’ means a ventilator was required, as part of COVID-19 patient care. ‘New_diabetes’ is diabetes diagnosed subsequent to the COVID-19 diagnosis, ascertained from the date of first evidence of diabetes using PHDC records.

Comorbidity data were provided for six comorbidities: Human Immunodeficiency Virus (HIV), chronic kidney disease (CKD), chronic obstructive pulmonary disease (COPD) or asthma, hypertension, diabetes mellitus (DM) and tuberculosis (TB). TB was further stratified into ‘TB-current’, a TB episode ongoing at time of COVID-19 diagnosis; and ‘TB-previously’, a TB episode occurring prior to COVID-19 diagnosis. Comorbidity episodes are inferred from a combination of facility visits, laboratory results and medications, are not clinician-validated for individual patients, and may have some margin of error. Data were not available for cardiovascular disease (CVD) episodes in this group. In brief: diabetes status was inferred in people who ever had an HBA1c of ≥ 6.5%, a 2-hr glucose of ≥ 11.1mmol/l after an oral glucose tolerance test, had been dispensed oral hypoglycaemic agents used exclusively for the management of diabetes or insulin, or had been assigned a diabetes ICD-10 diagnostic code [12]. For CKD, TB and HIV, laboratory results, specialist facility visits, and dispensed medications were used to define the nature of the condition. For hypertension and COPD or asthma, only dispensed medication was used to define the condition, without identifying patients with specific diagnoses who were untreated. Currently, CVD algorithms are under development and were not available for this study.

Pharmacy data from PHDC described the dispensing date for medicines. The association of medication on hospitalisation and mortality in PLWD diagnosed with COVID-19 was analysed from a subset of patients collecting medication from healthcare facilities linked to the electronic pharmacy records. Patient counts estimated the proportions of patients on various drugs in the year preceding COVID-19 diagnosis and post diagnosis, recognising that not all issued pharmacy drugs are always captured in the PHDC records. The medications selected and grouped were as follows: oral diabetes drugs (metformin, glimepiride), insulin (actrapid, Protaphane®, Actraphane®, Humulin N®, Humlin R®, Humulin 30/70®, insulin lispro, insulin aspart, insulin glargine, insulin detemir), statin (simvastatin, atorvastatin), angiotensin converting enzyme (ACE) inhibitor (enalapril), angiotensin receptor blocker (ARB) (losartan), steroids (dexamethasone, prednisone), hydrochlorothiazide (HCTZ) and anti-retroviral therapy (ART). These drugs were selected specific classes reflecting the formulary for the Western Cape Department of Health.

Laboratory data for all patients admitted to hospital for COVID-19 were analysed. The first available blood results in the period 2 days before and up to 5 days after the hospital admission for COVID-19, were considered as ‘‘admission investigations”.

4.3.3. Outcomes

We assessed the cumulative incidence of hospital admissions and deaths in PLWD diagnosed with COVID-19. For COVID-19 cases, the PHDC collates deaths data from hospital records, forensic pathology services, the National Institute for Communicable disease (NICD) notifications and death certification records.

Summary statistics were generated for the whole population and stratified by different sub-groups in the population. For continuous data, median and interquartile range were calculated and for grouped data, percentages were calculated. Multivariate logistic regression was used to estimate the effect on two outcomes, hospital admission and mortality and included all available co-variates.

4.3.4. Ethics

The study was approved by the University of Cape Town Faculty of Health Sciences Ethics Review Board (HREC Ref: 286/2020). As this study comprised anonymized and perturbed data, a waiver was granted for informed consent.

4.4. Results

4.4.1. Patient characteristics

Selection of the study population is described in Figure 1. Of approximately 4.0 million active patients aged ≥ 20 years in the PHDC database, 64 476 were diagnosed with COVID-19 by 15 July 2020, of whom 2993 (4.6%) died (Table 4.1).

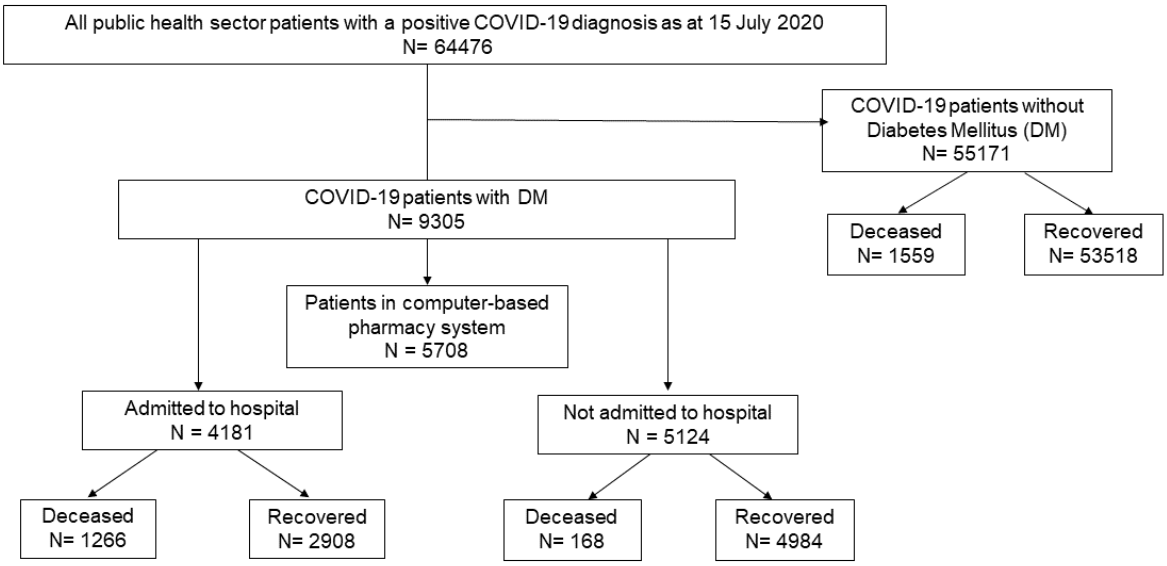

All public health sector patients with a positive COVID-19 diagnosis as at 15 July 2020
N= 64476
COVID-19 patients without Diabetes Mellitus (DM)
N= 55171
Deceased
N= 1559
Recovered
N= 53518
COVID-19 patients with DM
N= 9305
Patients in computer-based pharmacy system
N = 5708
Admitted to hospital
N = 4181
Not admitted to hospital
N = 5124
Deceased
N= 1266
Recovered
N= 2908
Deceased
N= 168
Recovered
N= 4984

Figure 4.1. Flow chart showing the selection of the study population from the PHDC routine health data.

4.4.1.1. The COVID-19 patient population

Population pyramids illustrate the youthful distribution of the Western Cape population, with 80% of the population under 50 years of age (Figure. 4.2A). For COVID-19 patients, a greater proportion (61.7%, n = 39 752) were women. Most patients (45.6%, n = 29 379) were 18 to 39 years old, with 70% of patients ≤ 50 years old. In this group, hypertension (19.6%, n = 12 623), diabetes (14.4%, n = 9305) and HIV infection (12.3%, n = 7933) were most prevalent co-morbidities. There were 16.9% (n = 10 887) hospitalised, 1.4% (n = 917) required intensive care, and 0.2% (n = 130) needed ventilation (Table 4.1).

Table 4.1. Characteristics of the Western Cape public health sector patients with COVID-19. The results have been grouped by Diabetes status (No Diabetes or Diabetes).

	ALL	**No Diabetes**	**Diabetes**
	N=64476	*N=55171*	*N=9305*
Sex:			
Female	39752 (61.7%)	34107 (61.8%)	5645 (60.7%)
Male	24669 (38.3%)	21012 (38.1%)	3657 (39.3%)
Age (years)	40.0 [30.0;52.0]	37.0 [29.0;49.0]	55.0 [46.0;63.0]
Age distribution:			
0-18	2654 (4.1%)	2635 (4.8%)	19 (0.2%)
18-39	29379 (45.6%)	28135 (51.0%)	1244 (13.4%)
40-49	13098 (20.3%)	11246 (20.4%)	1852 (19.9%)
50-59	10613 (16.5%)	7727 (14.0%)	2886 (31.0%)
60-69	5323 (8.3%)	3260 (5.9%)	2063 (22.2%)
70-79	2340 (3.6%)	1390 (2.5%)	950 (10.2%)
>=80	1069 (1.7%)	778 (1.4%)	291 (3.1%)
Outcome:			
Active	11 (<0.1%)	10 (<0.1%)	1 (<0.1%)
Died	2993 (4.6%)	1559 (2.8%)	1434 (15.4%)
Recovered	61374 (95.3%)	53518 (97.2%)	7856 (84.6%)
HIV	7933 (12.3%)	7022 (12.7%)	911 (9.8%)
TB current	791 (1.2%)	679 (1.2%)	112 (1.2%)
TB previously	3945 (6.1%)	3302 (6.0%)	643 (6.9%)
Asthma or COPD	4202 (6.5%)	2981 (5.4%)	1221 (13.1%)
Hypertension	12623 (19.6%)	7462 (13.5%)	5161 (55.5%)
CKD	1448 (2.2%)	596 (1.1%)	852 (9.2%)
Pregnant	958 (1.5%)	873 (1.6%)	85 (0.9%)
Hospital admission	10887 (16.9%)	6706 (12.2%)	4181 (44.9%)
ICU admission	917 (1.4%)	544 (1.0%)	373 (4.0%)
Ventilated	130 (0.2%)	78 (0.1%)	52 (0.6%)
New diabetes	1053 (11.3%)	0 (0.0%)	1053 (11.3%)

HIV, Human immunodeficiency virus; TB, Tuberculosis; COPD, Chronic obstructive pulmonary disease; CKD, chronic kidney disease; ICU, intensive care unit.

4.4.1.2. The diabetes/COVID-19 patient population

In 9305 PLWD with COVID-19, 11.3% (n = 1053) were newly diagnosed with DM during their COVID-19 episode. The 45–69-year-old age group had the most COVID-19 cases, with 66.5% of patients ≥ 50 years old and most patients in the 50–59 years category (31.0%; n = 2886). More women (60.7%, n = 5645) were diagnosed across all age groups (Figure. 2B). Hospital admissions appeared similar across genders, with the largest proportion of admissions for the 45–69-year age group (Figure. 2C). The distribution of COVID-19 deaths was similar, but men aged 55–69 years had the highest mortality (Figure. 2D). Compared to those without diabetes, a larger proportion of PLWD with COVID-19 were hospitalised (44.9%, c.f. 12.2%), admitted to ICU (4.0% c.f. 1.0%), and required ventilation (0.6% c.f. 0.1%) (Table 4.1).

4.4.2. Risk factors for admission and mortality in patients with COVID19

Logistic regression assessed the association of patient comorbidities and demographics with hospital admissions and mortality. For the total population of COVID-19 patients, current TB (OR:5.39, 95% CI: 4.61, 6.29), DM (OR:3.73, 95% CI: 3.53, 3.94), CKD (OR:1.87, 95% CI: 1.65, 2.10), COPD (OR:1.66, 95% CI: 1.54, 1.79), HIV infection (OR:1.64, 95% CI: 1.53, 1.75), male sex (OR:1.35, 95% CI: 1.29, 1.41), age per 5-year intervals (OR:1.18, 95% CI: 1.17, 1.19) were all associated with an increased risk for admission to hospital. Treated hypertension and previous TB were not associated with an increased risk for admission (Figure 4.3A).

Current TB (OR:4.68, 95% CI: 3.74, 5.82), DM (OR:3.01, 95% CI: 2.76, 3.28), HIV infection (OR:2.06, 95% CI: 1.82, 2.32), CKD (OR:1.82, 95% CI: 1.58, 2.09), male sex (OR:1.65, 95% CI: 1.52, 1.79), age per 5-year intervals (OR:1.42, 95% CI: 1.40, 1.44) and previous TB (OR:1.27, 95% CI: 1.10, 1.47) were associated with an increased risk for mortality. Treated hypertension (OR:0.91, 95% CI: 0.83, 0.99) was however, associated with a reduced mortality and COPD appeared to have a neutral effect on mortality (Figure 4.3B).

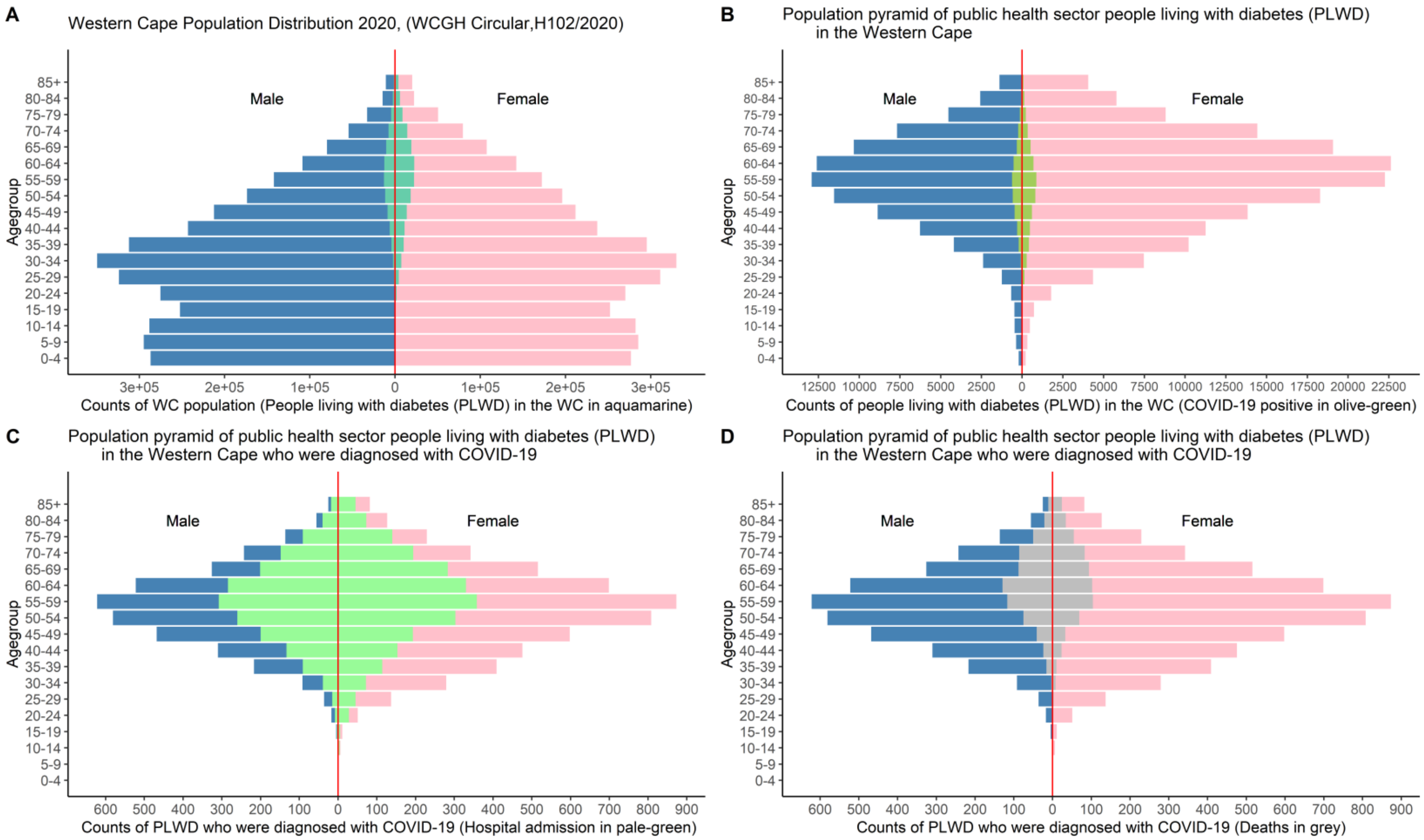


Figure 4.2. Population pyramids showing the distribution of people living with diabetes (PLWD) in the Western Cape (WC) population (A), PLWD with COVID-19 (B), PLWD with COVID-19 who got admitted into hospital for COVID-19 (C) and PLWD who died from COVID-19 (D).

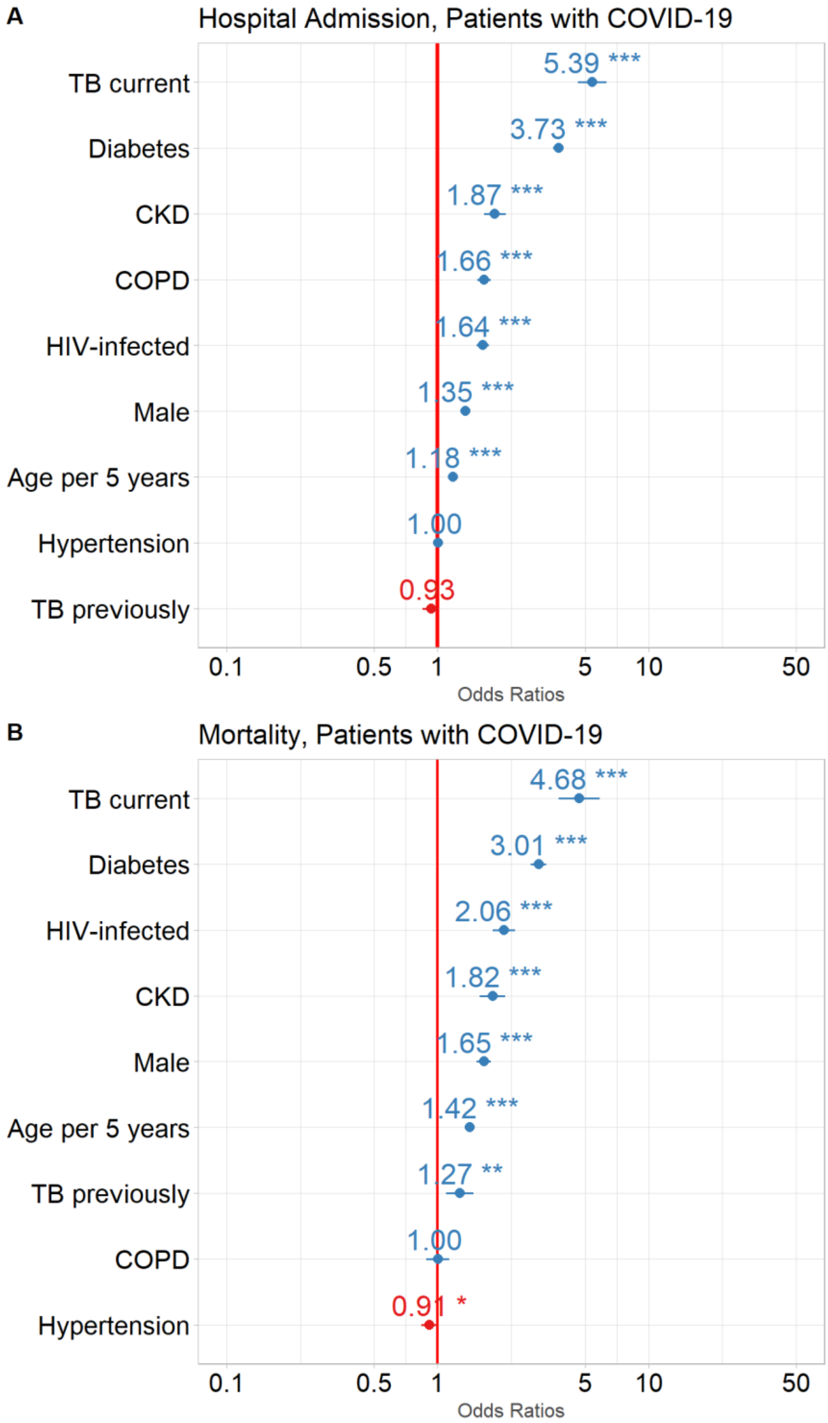


Figure 4.3. Impact of comorbidities and demographics on COVID-19 patient outcomes. Odds Ratios (circles) with 95% Confidence Intervals (horizontal lines) are shown for COVID19 patient outcomes: A. Admission to hospital, and B. Mortality (death). ***p<0.0001, **p<0.001, *p<0.01. TB, tuberculosis; CKD, chronic kidney disease; COPD, chronic obstructive pulmonary disease; HIV, human immunodeficiency virus.

4.4.3. Risk factors for hospitalisation and mortality in PLWD diagnosed with COVID-19

For PLWD who had COVID-19, HIV infection (OR:1.67, 95% CI: 1.44, 1.93), CKD (OR:1.60, 95% CI: 1.39, 1.87), current TB (OR:1.55, 95% CI: 1.05, 2.29), male sex (OR:1.41, 95% CI: 1.29, 1.54) and age per 5-year intervals (OR:1.15, 95% CI: 1.13, 1.17), were associated with an increased risk for hospitalisation. Treated hypertension (OR:0.73, 95% CI: 0.67, 0.79) was associated with a reduced risk for hospitalisation. Previous TB was not associated with hospitalisation (Figure 4.4A). CKD (OR:1.71, 95% CI: 1.44, 2.02), male sex (OR:1.70, 95% CI: 1.51, 1.92), HIV infection (OR:1.62, 95% CI: 1.32, 1.98), current TB (OR:1.59, 95% CI: 0.99, 2.50), previous TB (OR:1.54, 95% CI: 1.23, 1.90) and age per 5-year intervals (OR:1.33, 95% CI: 1.30, 1.37) were associated with an increased risk of mortality. Treated hypertension (OR:0.76, 95% CI: 0.67, 0.86) was also associated with a reduced risk for mortality and COPD was not associated with mortality (Figure 4.4B).

4.4.4. The effect of medication on hospitalisation and mortality in PLWD

Dispensing records for the preceding six months were available for 61.4% (n = 5708) of PLWD who were diagnosed with COVID-19 and who accessed their healthcare from a facility with a computer-based pharmacy system. Of these, 928 (16.3%) died and 4780 (83.7%) recovered (Supplementary Table 4.1). When comparing PLWD who were diagnosed with COVID19 and who accessed their healthcare from a facility with a computer-based pharmacy system to those PLWD who accessed their healthcare from a facility with no computer based pharmacy system it was noted that they were older [57.0 (48.0; 65.0) vs 52.0 (42.0; 61.0)] years old; 72.3% >50 years old vs 57.2% >50 years old] and were more likely to have comorbidities such as HIV infection, current TB, asthma/COPD, hypertension and CKD but had a similar outcome (Supplementary Table 4.2). Furthermore, in this cohort of patients accessing their healthcare from a facility with a computer based pharmacy system there were similar risk factors associated with hospitalisation and mortality when compared to the whole group (Figure 4.5A and 4.5B).

Use of insulin (OR:1.39, 95% CI: 1.24,1.57), was associated with an increased risk for hospitalisation whereas use of hydrochlorothiazide (OR:0.87, 95% CI: 0.77,0.97), a statin (OR:0.83, 95% CI: 0.72, 0.94) and metformin (OR:0.62, 95% CI: 0.55,0.71) were associated with a reduced risk for hospitalisation. Use of steroids, ARB, beta-blocker, ART, aspirin, ACE-I, or a sulphonylurea were not associated with hospitalisation (Figure 4.5A). Use of insulin (OR1.49, 95% CI: 1.27; 1.74), ARB (OR 1.34, 95% CI: 1.06; 1.70) and aspirin (OR1.24, 95% CI: 1.05; 1.46) were associated with an increased mortality whereas use of metformin (OR 0.77, 95% CI: 0.64; 0.92) was associated with a reduction in mortality. The use of steroids, beta-blocker, ACE-I, ART, sulphonylurea, statin and hydrochlorothiazide were not associated with mortality (Figure 4.5B).

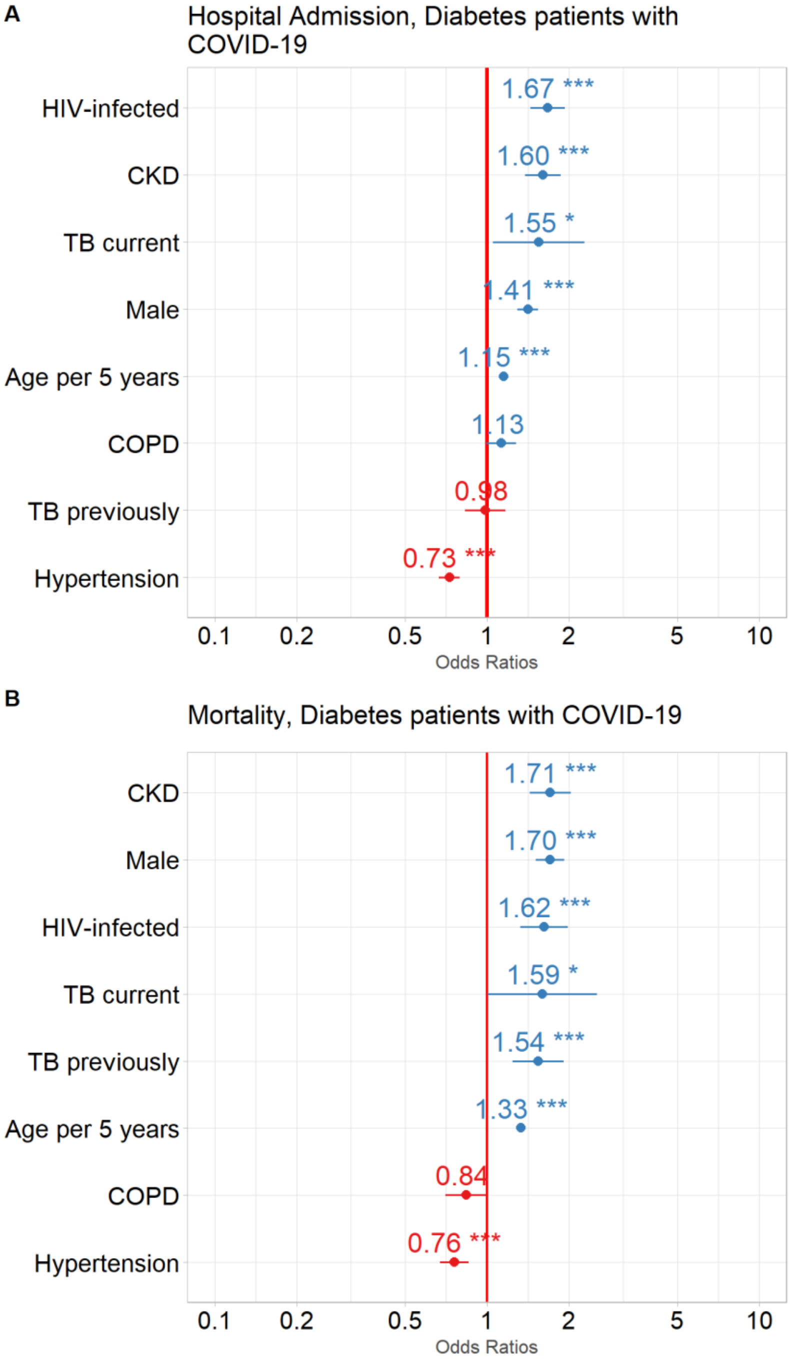


Figure 4.4. Impact of comorbidities and demographics on outcomes in COVID-19 patients with DM. Odds Ratios (circles) with 95% Confidence Intervals (horizontal lines) are shown for COVID-19 patient outcomes: A. Admission to hospital, and B. Mortality (death). ***p<0.0001, **p<0.001, *p<0.01. HIV, human immunodeficiency virus; CKD, chronic kidney disease; TB, tuberculosis; COPD, chronic obstructive pulmonary disease.

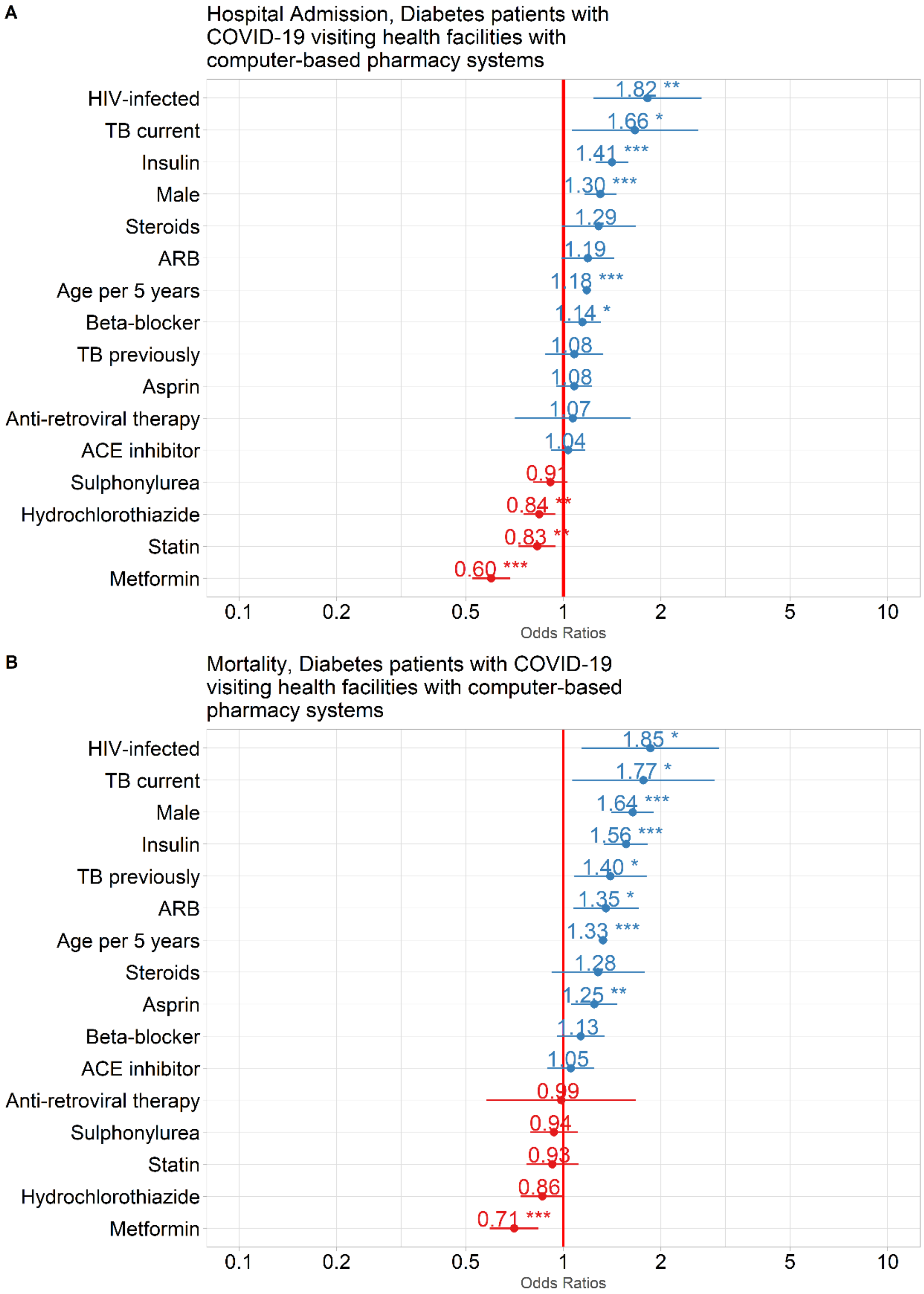


Figure 4.5. Impact of common comorbidities, medications dispensed in the 6 months prior to COVID-19 diagnosis and demographic factors to the outcomes of COVID-19 patients with DM. Odds Ratios (circles) with 95% Confidence Intervals (horizontal lines) are shown for COVID-19 patient outcomes: A. Admission to hospital, and B. Mortality (death). ***p<0.0001, **p<0.001, *p<0.01. HIV, human immunodeficiency virus; TB, tuberculosis; CKD, chronic kidney disease; COPD, chronic obstructive pulmonary disease; ARB, angiotensin receptor blocker; ACE, angiotensin converting enzyme.

4.4.5. Risk factors for mortality in PLWD hospitalised with COVID-19

Detailed laboratory investigations were available for 3664 PLWD who were hospitalised, shown in detail in Supplementary Table 4.3. All stages of CKD were associated with an increased mortality [stage 5 (OR:5.53, 95% CI: 3.60, 8.60), stage 4 (OR:3.44, 95% CI: 2.39, 4.98), stage 3B (OR:3.68, 95% CI: 2.63, 5.16), stage 3A (OR:1.45, 95% CI: 1.08, 1.93)]. Other factors associated with an increased mortality were male sex (OR:1.54, 95% CI: 1.26, 1.89) and age per 5-year intervals (OR:1.21, 95% CI: 1.15, 1.26) (Figure 4.6).

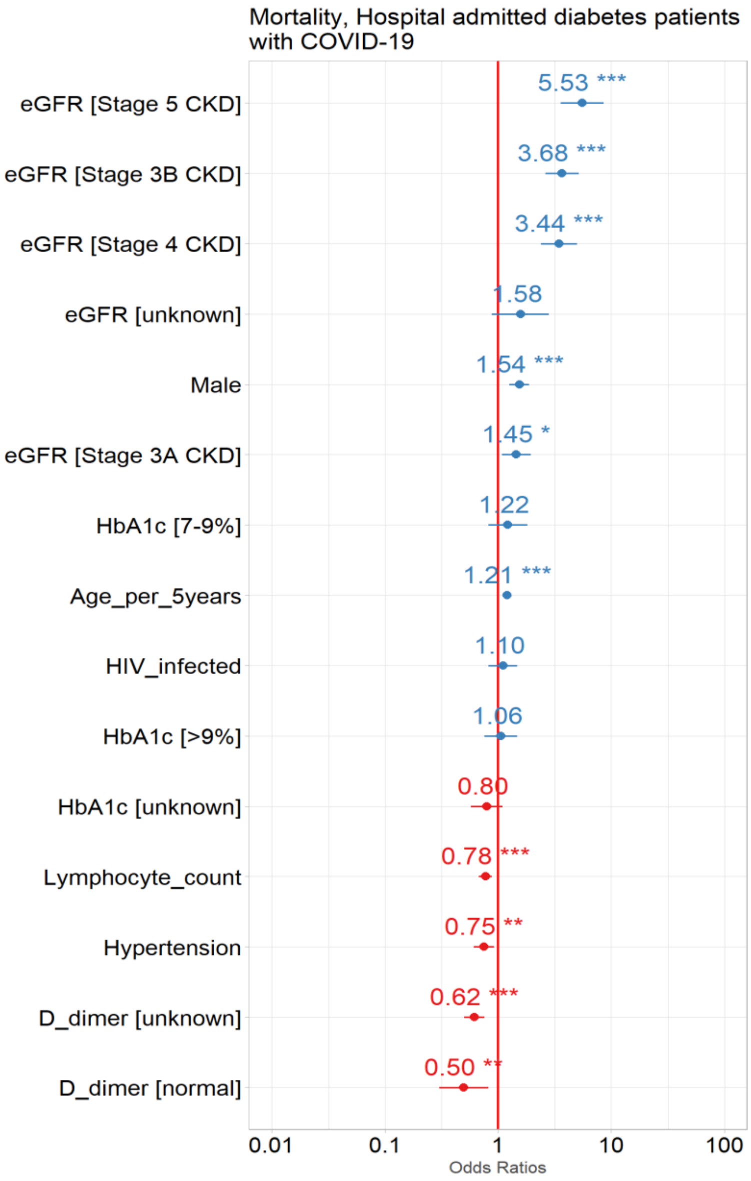


Figure 4.6. Impact of common comorbidities and laboratory results at admission, as well as demographic factors to mortality in hospital admitted diabetes patients with COVID-19. Odds Ratios (circles) with 95% Confidence Intervals (horizontal lines) are shown for COVID-19

patient outcomes. ***p<0.0001, **p<0.001, *p<0.01. eGFR, estimated glomerular filtration rate; HIV, human immunodeficiency virus.

Treated hypertension (OR:0.77, 95% CI: 0.61, 0.92), a normal lymphocyte count (OR:0.76, 95% CI: 0.68, 0.88) and an unknown (OR:0.62, 95% CI: 0.50, 0.76) or normal (OR:0.50, 95% CI: 0.30, 0.80) d-dimer were associated with a reduced risk for mortality. No level of HbA1c was associated with mortality (Figure 4.6).

4.5. Discussion

To our knowledge, this is the largest study of patients with COVID-19 from Africa and the Southern Hemisphere, which describes the association of diabetes with severe COVID-19 (hospitalisation and mortality), adjusting for key confounding factors. The key findings were that increasing age, male sex, diabetes, current tuberculosis, HIV infection and chronic kidney disease were significantly associated with admission to hospital and mortality. In PLWD, HIV-infection, chronic kidney disease, current tuberculosis, male sex and increasing age were significantly associated with admission to hospital and mortality. Use of metformin was associated with a reduced risk of hospitalisation and mortality in PLWD.

Our results show an increased risk of hospital admission, ICU admission and death in people diagnosed with COVID-19 who have diabetes, with half of all COVID-19 deaths occurring in people with diabetes. After adjusting for various variables such as age, sex and co-morbid disease, including HIV status, PLWD had almost four times the risk for hospitalisation and three times the risk of death relative to people without diabetes. This larger study also confirms earlier findings from an analysis performed before 1 June 2020, of COVID-19 death in the general population, being associated with age, male sex, chronic kidney disease, and in people with active tuberculosis and HIV (Western Cape Department of Health in collaboration with the National Institute for Communicable Diseases, 2021).

This high risk for hospitalisation and death for PLWD is well recognised in other series. Metanalyses of studies including small numbers of patients with COVID-19 have reported a mortality risk of between 1.9-3.5 for patients with diabetes (Apicella et al., 2020) and the largest cohort study of PLWD in primary care reported a risk for mortality of 2.03 for patients with T2DM (Barron et al., 2020). In this cohort mortality was high with 15.4% of PLWD and COVID-19 having died, 3.3% of those not hospitalised and 30.3% of those admitted to hospital. Early data from Wuhan reported a mortality rate of 20% for hospitalised patients with COVID-19 and diabetes (Shi et al., 2020). This mortality rate for hospitalised patients is similar to a retrospective cohort analysis of 1126 patients with diabetes hospitalised with COVID-19 at a large academic medical centre in New York City, where the mortality rate was 33.1%, but in patients with a mean age almost 10 years older than in this cohort (Agarwal et al., 2020). Updated results from the French CORONADO Study of 2796 PLWD (mean age 69.7) who were hospitalised reported a mortality rate of 11.2% within 7 days, which continued to increase to 20.6% by 28 days (Wargny et al., 2021). We can only speculate about factors that account for such a high mortality in our setting, even in relatively young patients, which includes social determinants such as poor access to care, general poor standards of chronic care for diabetes such as glycaemic control and management of complications, higher threshold for admission and reduced access to ICU beds once admitted (Govender et al., 2012). The high background of infectious diseases in the population may also contribute to mortality.

Amongst PLWD we corroborated well-described associations for death including age, male sex, and CKD (Apicella et al., 2020; Holman et al., 2020). In this population, 9.8% were infected with HIV and 1.2% had active tuberculosis, with concurrent HIV and tuberculosis exhibiting more severe disease, as evidenced by higher hospitalisation and death rates. PLWD with treated hypertension had a reduced risk of mortality in our dataset, acknowledging that the hypertension episode is inferred from prescribed medications, and the impact of untreated hypertension on outcomes could not be assessed. Although hypertension has frequently been found to be associated with poor outcomes of COVID-19 in other studies, we could not find an association and even described a small but significant reduction in risk of poor outcomes in our cohort. McGurnaghan et al (2021), also failed to demonstrate an association between HPT or HPT treatment and worse outcomes in PLWD in Scotland (McGurnaghan et al., 2021). In our dataset, a hypertension episode is defined only on the basis of patients receiving hydrochlorothiazide dispensed as an antihypertensive and blood pressure observational data are not available for analysis, meaning that undiagnosed/untreated hypertension is not documented in the current dataset. Given that Berry et al (2017), described that 48.7% of South Africans with hypertension are undiagnosed (Berry et al., 2017), it is plausible that patients who are receiving medication for hypertension have better outcomes than the many patients in the cohort who are likely to have undiagnosed and/or untreated hypertension, and the population of patients that are diagnosed with hypertension and dispensed medication may be different from those in other series. In the subset of PLWD for whom full prescription data were available for the 6 months preceding the diagnosis of COVID-19, use of ARB, insulin and aspirin had a moderate independent association with mortality, yet metformin was protective. Episode data for cardiovascular disease, as a risk factor were not available for this cohort, but dispensing records for aspirin, statin, betablocker, ACE and ARB were all associated with worse outcomes in the cohort. We postulate that these medications may be a surrogate marker for cardiovascular disease, potentially explaining the higher risk of poor outcomes in patients receiving these medications. An early report from Wuhan, China, also suggested that metformin was associated with a decreased mortality in hospitalized COVID-19 patients (Luo et al., 2020). Moreover, metformin was found to be associated with reduced risk of early death in the French CORONADO study (Lalau et al., 2021) and with decreased mortality in women with COVID-19, based on a United Health data analysis (Bramante et al., 2021). Three metanalyses have (Hariyanto & Kurniawan, 2020; Kow & Hasan, 2021; Lukito et al., 2020) corroborated these findings. It is possible that metformin use, may reflect PLWD with less advanced disease and fewer complications, such as chronic kidney disease in which it is contraindicated or possibly, patients who are more adherent to prescribed therapy. A recent paper using propensity scoring to adjust for such con founders failed to show an association of metformin use on susceptibility or outcome (Wang et al., 2021). Although we show, as have others, that use of insulin is associated with worse outcomes, this most likely simply reflects a group of PLWD of longer disease duration and thus more vascular and other complications.

We did not demonstrate an association with HbA1c results and poorer outcomes in PLWD. Similarly, the Italian CORONADO Study and the population-based study from Scotland also failed to demonstrate an association between HbA1c and mortality (Cariou et al., 2021; McGurnaghan et al., 2021). However, there are studies that have shown an association between poor outcome and worse glycaemic control. In some of these population studies PLWD were compared to a 'non-diabetic'' population [9], but in other studies where increments in HbA1c levels were compared exclusively in PLWD an association with poorer outcomes was shown (Holman et al., 2020). Our data should, however, be considered with caution, as only a small proportion (31.5%) of PLWD had an HbA1c analysed in the preceding year. The reasons for this could be due to poor linkage to care, and also may reflect missing data due to patients opting for either public or private health care depending on their changing financial and employment circumstances. Private healthcare data in our system are not available, limiting analysis of this subset.

Strengths of the study include the large study size, with near complete ascertainment of outcome data, laboratory confirmed SARS-CoV-2 diagnosis in all COVID-19 cases, the inclusion of hospitalised and non-hospitalised cases and deaths, as well as modelling the independent association of diabetes with death. Patients newly diagnosed with diabetes during testing or admission, based on a diagnostic HbA1c were also included. Limitations include that the classification cascade used to identify somebody as having diabetes using the routine health data in the PHDC is still in the process of formal validation, but, if anything, may provide an underestimate of actual diabetes prevalence. Also, given the predominantly administrative nature of data capture, important comorbidities, lack of data on other potential risk factors including socio-economic status, lack of observational clinical data such as blood pressure, smoking and body mass index (BMI) limit the included comorbidities. In addition, data relating to some biochemical variables may be incomplete due to potential public/private health sector patient mobility, for example HbA1c data. Admission criteria for ICU and data capturing in the various facilities in the Western Cape is variable and possibly incomplete. The missing data for some comorbidities for example, cardiovascular disease, along with the absence of other potential confounders may result in large residual confounding in the associations described.

In conclusion, there is convincing evidence from many large population-based studies that PLWD are at greater risk of severe COVID-19 disease (hospital and ICU admissions) and of death than those without diabetes. This study adds to the body of evidence from a low and middle-income country, where the burden of DM affects younger people, compared to high income countries where older people are predominantly affected. We show that the population with diabetes is at particularly high risk, possibly due to poorer access to optimal care for diabetes. We also show that concurrent HIV infection and DM are associated with more severe disease and that metformin use in particular, is associated with a reduced mortality. These findings are of major public health importance, which raise the question of how to ameliorate the high-risk burden of PLWD in COVID-19. It is incumbent upon us as healthcare providers to offer education and close monitoring of risk in PLWD. It is too premature to recommend widespread use of metformin. Additional interventions may include home oxygen saturation monitoring, ensuring adequate glycaemic control, early identification of deterioration in symptoms with rapid access to hospital admission and consideration for pre-emptive admissions to hospital for those PLWD who are deemed to be at very high risk of severe COVID-19.

4.6. References

Agarwal, S., Schechter, C., Southern, W., Crandall, J. P., & Tomer, Y. (2020). Preadmission Diabetes-Specific Risk Factors for Mortality in Hospitalized Patients With Diabetes and Coronavirus Disease 2019. *Diabetes Care*, *43*(10), 2339–2344. https://doi.org/10.2337/dc20-1543

Apicella, M., Campopiano, M. C., Mantuano, M., Mazoni, L., Coppelli, A., & Del Prato, S. (2020). COVID-19 in people with diabetes: Understanding the reasons for worse outcomes. *The Lancet Diabetes & Endocrinology*, *8*(9), 782–792. https://doi.org/10.1016/S2213-8587(20)30238-2

Barron, E., Bakhai, C., Kar, P., Weaver, A., Bradley, D., Ismail, H., Knighton, P., Holman, N., Khunti, K., Sattar, N., Wareham, N. J., Young, B., & Valabhji, J. (2020). Associations of type 1 and type 2 diabetes with COVID-19-related mortality in England: A whole-population study. *The Lancet Diabetes & Endocrinology*, *8*(10), 813–822. https://doi.org/10.1016/S2213-8587(20)30272-2

Bello-Chavolla, O. Y., Bahena-López, J. P., Antonio-Villa, N. E., Vargas-Vázquez, A., González-Díaz, A., Márquez-Salinas, A., Fermín-Martínez, C. A., Naveja, J. J., & Aguilar-Salinas, C. A. (2020). Predicting Mortality Due to SARS-CoV-2: A Mechanistic Score Relating Obesity and Diabetes to COVID-19 Outcomes in Mexico. *The Journal of Clinical Endocrinology & Metabolism*, *105*(8), 2752–2761. https://doi.org/10.1210/clinem/dgaa346

Berry, K. M., Parker, W., Mchiza, Z. J., Sewpaul, R., Labadarios, D., Rosen, S., & Stokes, A. (2017). Quantifying unmet need for hypertension care in South Africa through a care cascade: Evidence from the SANHANES, 2011-2012. *BMJ Global Health*, *2*(3), e000348. https://doi.org/10.1136/bmjgh-2017-000348

Boulle, A., Heekes, A., Tiffin, N., Smith, M., Mutemaringa, T., Zinyakatira, N., Phelanyane, F., Pienaar, C., Buddiga, K., Coetzee, E., Rooyen, R. van, Dyers, R., Fredericks, N., Loff, A., Shand, L., Moodley, M., Vega, I. de, & Vallabhjee, K. (2019). Data Centre Profile: The Provincial Health Data Centre of the Western Cape Province, South Africa. *International Journal of Population Data Science*, *4*(2), Article 2. https://doi.org/10.23889/ijpds.v4i2.1143

Bramante, C. T., Ingraham, N. E., Murray, T. A., Marmor, S., Hovertsen, S., Gronski, J., McNeil, C., Feng, R., Guzman, G., Abdelwahab, N., King, S., Tamariz, L., Meehan, T., Pendleton, K. M., Benson, B., Vojta, D., & Tignanelli, C. J. (2021). Metformin and risk of mortality in patients hospitalised with COVID-19: A retrospective cohort analysis. *The Lancet Healthy Longevity*, *2*(1), e34–e41. https://doi.org/10.1016/S2666-7568(20)30033-7

Cariou, B., Hadjadj, S., Wargny, M., Pichelin, M., Al-Salameh, A., & Allix, I. (2020). Phenotypic characteristics and prognosis of inpatients with COVID-19 and diabetes: The CORONADO study. *Phenotypic characteristics and prognosis of inpatients with COVID-19 and diabetes: the CORONADO study*, 1–16. Scopus.

Cariou, B., Pichelin, M., Goronflot, T., Gonfroy, C., Marre, M., Raffaitin-Cardin, C., Thivolet, C., Wargny, M., Hadjadj, S., & Gourdy, P. (2021). Phenotypic characteristics and prognosis of newly diagnosed diabetes in hospitalized patients with COVID-19: Results from the CORONADO study. *Diabetes Research and Clinical Practice*, *175*, 108695. https://doi.org/10.1016/j.diabres.2021.108695

COVID Live—Coronavirus Statistics—Worldometer. (n.d.). Retrieved January 7, 2023, from https://www.worldometers.info/coronavirus/

COVID-19: Perspectives from people with diabetes. (2020). *Diabetes Research and Clinical Practice*, *163*, 108201. https://doi.org/10.1016/j.diabres.2020.108201

Govender, I., Ehrlich, R., Van Vuuren, U., De Vries, E., Namane, M., De Sa, A., Murie, K., Schlemmer, A., Govender, S., Isaacs, A., & Martell, R. (2012). Clinical audit of diabetes management can improve the quality of care in a resource-limited primary care setting. *International Journal for Quality in Health Care*, *24*(6), 612–618. https://doi.org/10.1093/intqhc/mzs063

Hariyanto, T. I., & Kurniawan, A. (2020). Metformin use is associated with reduced mortality rate from coronavirus disease 2019 (COVID-19) infection. *Obesity Medicine*, *19*, 100290. https://doi.org/10.1016/j.obmed.2020.100290

Holman, N., Knighton, P., Kar, P., O’Keefe, J., Curley, M., Weaver, A., Barron, E., Bakhai, C., Khunti, K., Wareham, N. J., Sattar, N., Young, B., & Valabhji, J. (2020). Risk factors for COVID-19-related mortality in people with type 1 and type 2 diabetes in

England: A population-based cohort study. *The Lancet Diabetes & Endocrinology*, *8*(10), 823–833. https://doi.org/10.1016/S2213-8587(20)30271-0

Kow, C. S., & Hasan, S. S. (2021). Mortality risk with preadmission metformin use in patients with COVID-19 and diabetes: A meta-analysis. *Journal of Medical Virology*, *93*(2), 695–697. https://doi.org/10.1002/jmv.26498

Lalau, J.-D., Al-Salameh, A., Hadjadj, S., Goronflot, T., Wiernsperger, N., Pichelin, M., Allix, I., Amadou, C., Bourron, O., Duriez, T., Gautier, J.-F., Dutour, A., Gonfroy, C., Gouet, D., Joubert, M., Julier, I., Larger, E., Marchand, L., Marre, M., … Cariou, B. (2021). Metformin use is associated with a reduced risk of mortality in patients with diabetes hospitalised for COVID-19. *Diabetes & Metabolism*, *47*(5), 101216. https://doi.org/10.1016/j.diabet.2020.101216

Lukito, A. A., Pranata, R., Henrina, J., Lim, M. A., Lawrensia, S., & Suastika, K. (2020). The Effect of Metformin Consumption on Mortality in Hospitalized COVID-19 patients: A systematic review and meta-analysis. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, *14*(6), 2177–2183. https://doi.org/10.1016/j.dsx.2020.11.006

Luo, P., Qiu, L., Liu, Y., Liu, X., Zheng, J., Xue, H., Liu, W., Liu, D., & Li, J. (2020). Metformin Treatment Was Associated with Decreased Mortality in COVID-19 Patients with Diabetes in a Retrospective Analysis. *The American Journal of Tropical Medicine and Hygiene*, *103*(1), 69–72. https://doi.org/10.4269/ajtmh.20-0375

Mancia, G., Rea, F., Ludergnani, M., Apolone, G., & Corrao, G. (2020). Renin–Angiotensin–Aldosterone System Blockers and the Risk of Covid-19. *New England Journal of Medicine*, *382*(25), 2431–2440. https://doi.org/10.1056/NEJMoa2006923

Martins-Filho, P. R., Araújo, A. A. de S., Pereira, L. X., Quintans-Júnior, L. J., Barboza, W. de S., Cavalcante, T. F., Souza, M. F. de, Góes, M. A. de O., & Santos, V. S. (2021). Factors Associated with Mortality among Hospitalized Patients with COVID-19: A Retrospective Cohort Study. *The American Journal of Tropical Medicine and Hygiene*, *104*(1), 103–105. https://doi.org/10.4269/ajtmh.20-1170

McGurnaghan, S. J., Weir, A., Bishop, J., Kennedy, S., Blackbourn, L. A. K., McAllister, D. A., Hutchinson, S., Caparrotta, T. M., Mellor, J., Jeyam, A., O'Reilly, J. E., Wild, S. H., Hatam, S., Höhn, A., Colombo, M., Robertson, C., Lone, N., Murray, J., Butterly, E., … McCoubrey, J. (2021). Risks of and risk factors for COVID-19 disease in people with diabetes: A cohort study of the total population of Scotland. *The Lancet Diabetes & Endocrinology*, *9*(2), 82–93. https://doi.org/10.1016/S2213-8587(20)30405-8

Shi, Q., Zhang, X., Jiang, F., Zhang, X., Hu, N., Bimu, C., Feng, J., Yan, S., Guan, Y., Xu, D., He, G., Chen, C., Xiong, X., Liu, L., Li, H., Tao, J., Peng, Z., & Wang, W. (2020). Clinical Characteristics and Risk Factors for Mortality of COVID-19 Patients With Diabetes in Wuhan, China: A Two-Center, Retrospective Study. *Diabetes Care*, *43*(7), 1382–1391. https://doi.org/10.2337/dc20-0598

Wang, J., Cooper, J. M., Gokhale, K., Acosta-Mena, D., Dhalla, S., Byne, N., Chandan, J. S., Anand, A., Okoth, K., Subramanian, A., Bangash, M. N., Jackson, T., Zemedikun, D., Taverner, T., Hanif, W., Ghosh, S., Narendran, P., Toulis, K. A., Tahrani, A. A., … Nirantharakumar, K. (2021). Association of Metformin with Susceptibility to COVID-19 in People with Type 2 Diabetes. *The Journal of Clinical Endocrinology & Metabolism*, *106*(5), 1255–1268. https://doi.org/10.1210/clinem/dgab067

Wargny, M., Potier, L., Gourdy, P., Pichelin, M., Amadou, C., Benhamou, P.-Y., Bonnet, J.-B., Bordier, L., Bourron, O., Chaumeil, C., Chevalier, N., Darmon, P., Delenne, B., Demarsy, D., Dumas, M., Dupuy, O., Flaus-Furmaniuk, A., Gautier, J.-F., Guedj, A.-M., … for the CORONADO investigators. (2021). Predictors of hospital discharge and mortality in patients with diabetes and COVID-19: Updated results from the nationwide CORONADO study. *Diabetologia*, *64*(4), 778–794. https://doi.org/10.1007/s00125-020-05351-w

Western Cape Department of Health in collaboration with the National Institute for Communicable Diseases, S. A. (2021). Risk Factors for Coronavirus Disease 2019 (COVID-19) Death in a Population Cohort Study from the Western Cape Province, South Africa. *Clinical Infectious Diseases*, *73*(7), e2005–e2015. https://doi.org/10.1093/cid/ciaa1198

Williamson, E. J., Walker, A. J., Bhaskaran, K., Bacon, S., Bates, C., Morton, C. E., Curtis, H. J., Mehrkar, A., Evans, D., Inglesby, P., Cockburn, J., McDonald, H. I., MacKenna, B., Tomlinson, L., Douglas, I. J., Rentsch, C. T., Mathur, R., Wong, A. Y. S., Grieve, R., … Goldacre, B. (2020). Factors associated with COVID-19-related death using OpenSAFELY. *Nature*, *584*(7821), Article 7821. https://doi.org/10.1038/s41586-020-2521-4

5. Chapter 5: Routine health data describe adherence and persistence patterns for oral diabetes medication for a virtual cohort in the Khayelitsha sub-district of Cape Town, South Africa

Relevance of chapter in thesis

The management of T2DM includes the use of long-term anti-diabetes medication however, numerous research has shown that PLWD are not adherent to their treatment. Data on adherence in PLWD in sub-Saharan Africa has largely been limited to cross-sectional data using patient self-reporting. In this chapter we explore methods to use longitudinal routine health data to describe adherence and persistence patterns for diabetes medication use. We then use these methods to better understand adherence and persistence patterns to metformin in a virtual cohort of people with T2DM who are starting their diabetes treatment in the Western Cape Province of South Africa. The methods optimised in this chapter further highlighted the how longitudinal routine health data can be modelled to describe patient phenotypes at both an individual and population level.

5.1. Abstract

Introduction

Type 2 diabetes mellitus (T2DM) is managed using a combination of lifestyle modifications and antidiabetic drugs, with the aim of achieving glycaemic control. Studies have shown that people with T2DM who are on treatment often fail to reach glycaemic control. Given the significance of adherence in achieving optimal glycaemic control, and because management of diabetes with drugs is a lifelong process, it is important to understand adherence through the analysis of longitudinal medications data.

Methods

Using retrospective routine health data and recorded metformin dispensing episodes as a proxy for medication use, we describe longitudinal persistence and adherence to oral diabetes medication in a virtual cohort of people with diabetes (PLWD) in the Khayelitsha subdistrict in the Western Cape Province, South Africa. Adherence was measured in 120-day sliding windows in a two-year period and used to estimate metformin adherence trajectories. Multinomial logistic regression was used to identify factors which influence metformin adherence trajectories for individuals in the study population.

Results

An analysis of the pharmacy dispensing records showed that the study participants had varying medication refill patterns. While some refilled their prescriptions consistently others had gaps in treatment which resulted in periods of non-persistence and multiple treatment episodes across the two years assayed which ranged in number from one to as high as five in some participants. There was a general trend of decreasing adherence over time and across all sliding windows in the two-year observation window, and by the end of the two-year period only 25% of the study population achieved medication adherence (>= 80% adherence). Four adherence trajectories; ‘low adherence gradual decline (A), ‘high adherence rapid decline’ (B), ‘low adherence gradual increase (C) and ‘adherent’ (D) were identified. Only trajectory D represented participants who were adherent at treatment start and were still adherent at the end of two years. Taking HIV antiretroviral treatment before or concurrently with diabetes treatment and taking metformin in combination with sulphonylurea and/or insulin were all associated with a participant having long-term adherence (trajectory D). Increasing participant age at the start of treatment was associated with long-term non-adherence (trajectory A and B) while participant sex did not influence adherence trajectory.

Conclusion

Routine data shows real life medication implementation patterns which might not be seen under controlled study conditions. The findings from this study illustrate the utility of these data in describing longitudinal adherence patterns at both an individual and population level.

5.2. Introduction

Diabetes is one of the fastest growing global health threats with an estimated 1 in 10 people currently living with the disease(*IDF Diabetes Atlas 10th Edition*, n.d.). Type 2 diabetes mellitus (T2DM) is the most prevalent form of diabetes accounting for more than 90% of all cases (Campos, 2012; Fowler, 2008; Rask-Madsen & King, 2013; Zheng et al., 2018). Chronic hyperglycaemia increases the risk of developing micro- and macro-vascular complications which are associated with increased morbidity and mortality in people living with diabetes (PLWD) (Campos, 2012; Fowler, 2008; Rask-Madsen & King, 2013). To minimise the risk of developing complications, T2DM is managed using a combination of lifestyle modifications and antidiabetic medications with the aim of achieving glycaemic control - established as a glycated haemoglobin (HbA1c) of less than 7%("Intensive Blood-Glucose Control with Sulphonylureas or Insulin Compared with Conventional Treatment and Risk of Complications in Patients with Type 2 Diabetes (UKPDS 33)," 1998). Antidiabetic medications can effectively reduce hyperglycaemia and improve glycaemic control(Aikens & Piette, 2013; Chepulis et al., 2020; Cohen et al., 2010; "Intensive Blood-Glucose Control with Sulphonylureas or Insulin Compared with Conventional Treatment and Risk of Complications in Patients with Type 2 Diabetes (UKPDS 33)," 1998; Raum et al., 2012). Studies have shown, however, that PLWD who are on treatment often fail to reach glycaemic targets (Afroz et al., 2019; Camara et al., 2015; Erasmus et al., 1999; Govender et al., 2017; Lin et al., 2017; Musenge et al., 2015; Pinchevsky et al., 2015; Polonsky & Henry, 2016; Raum et al., 2012). While there are several factors that can contribute to lack of glycaemic control including the natural progression of the disease, one of the key factors is non-adherence to diabetes treatment (García-Pérez et al., 2013; Khunti et al., 2019; Polonsky & Henry, 2016): treatment non-adherence has been shown to negatively impact treatment efficacy and lead to increased morbidity and mortality in PLWD (Evans et al., n.d.; Ho et al., 2006; Lin et al., 2017).

Sub-Saharan Africa (SSA) currently has the lowest prevalence of diabetes in the world, but, has the highest diabetes-related morbidity and mortality in people under 60 years (*IDF Diabetes Atlas 10th Edition*, n.d.). Non-adherence to diabetes treatment has also been widely studied in SSA, but most of the data have been from cross sectional studies. These studies have reported varying trends of adherence ranging from as low as 25% to as high as 75% and identified several factors including sex, socio-economic status, age, high cost of medication, comorbidities, pill burden, medication availability, drug side-effects and being asymptomatic as having an influence on adherence to antidiabetic medication PLWD (Ali et al., 2017; Demoz et al., 2020; Kretchy et al., 2020; Rwegerera, 2014; Shilubane & Cur, 2010; Waari et al., 2018). While these studies have provided valuable insight about adherence to diabetes treatment PLWD in SSA, they are limited in that they rely on patient self-reporting, a subjective measure that is prone to bias and only estimates adherence at a single time point. Management of diabetes is a lifelong process, and an individual's medication adherence patterns can change over time, therefore, it is important to estimate adherence using longitudinal data (Egede et al., 2014; Lo-Ciganic et al., 2016).

There is a paucity of data on longitudinal adherence patterns to diabetes medication PLWD in SSA. This is worrying in a region that is predicted to soon have an exponential increase in diabetes prevalence (*IDF Diabetes Atlas 10th Edition*, n.d.). Given the significance of adherence in achieving optimal glycaemic control, estimating longitudinal adherence can better elucidate temporal adherence patterns which can be more informative and useful for improved patient care and targeted interventions (Egede et al., 2014; Lo-Ciganic et al., 2016). In the current study we explore methods to use longitudinal routine health data to describe adherence and persistence patterns for diabetes medication use. We then use these methods to better understand adherence and persistence patterns to metformin in a virtual cohort of people with T2DM who are starting their diabetes treatment in the Western Cape Province of South Africa.

5.3. Methods

5.3.1. Study population

The study population was selected from a pre-existing dataset of public health care seekers in the Khayelitsha sub-district in the Western Cape Province. The Khayelitsha subdistrict is a high-density urban area with generally poor socioeconomic conditions and a high burden of TB and HIV. The study population in the pre-existing dataset were selected from the Western Cape population as represented in the PHDC, a health information exchange containing routine health data for about 7 million healthcare clients, collated daily from multiple electronic health data sources in the Western Cape Province, South Africa(Boulle et al., 2019). Inclusion criteria in the pre-existing dataset were: (1) Having attended at least one Government Health Facility in the Khayelitsha sub-district in the Western Cape, South Africa, in the period 1 January 2016 to 31 December 2017, and (2) aged 18 or older by December 2017.

In the current study, inclusion criteria were: (1) a diagnosis of diabetes inferred from PHDC records using listed disease evidences of at least one glycated haemoglobin (HbA1c) value greater than or equal to 6.5%("Use of Glycated Haemoglobin (HbA1c) in the Diagnosis of Diabetes Mellitus," 2011), fasting glucose results, and/or dispensed diabetes drugs, and (2) diabetes medication dispensed at health facilities linked to electronic routine pharmacy records. Exclusion criteria were: (1) no recorded dispensing of metformin in the study observation window, (2) diabetes ascertainment at less than 18 years of age and diabetes treatment using insulin only, used as a proxy for early onset Type 1 Diabetes, (3) diabetes ascertainment occurring during pregnancy, used as a proxy for gestational diabetes, (4) diabetes treatment start after 31st March 2018 as there would be insufficient data for a two-year follow-up and (5) diabetes treatment start before 1st January 2011 as there would be insufficient data in the PHDC records.

The study had a two-year window for observing medication dispensing patterns for PLWD in the Khayelitsha sub-district. Diabetes treatment start date was defined as day 0 in the observation window and all individuals were followed up for two-years from their diabetes treatment start date. The study analysed retrospective routine health data and used recorded medication dispensing episodes as a proxy for medication use. Since metformin is a first-line drug for treating type 2 diabetes ("Effect of Intensive Blood-Glucose Control with Metformin on Complications in Overweight Patients with Type 2 Diabetes (UKPDS 34)," 1998; *Guidelines*, n.d.; Zaccardi et al., 2020), metformin dispensing was used as a proxy for diabetes treatment.

Retrospective PHDC data for 10541 individuals who were recorded as having started diabetes treatment on metformin were analysed together with population demographics assessed as of 31 December 2017, using descriptive statistics. A diagnosis is inferred by the PHDC using laboratory and pharmacy data and is not a clinical diagnosis made during a consultation, so is referred to as 'ascertainment' to make this distinction. The process of episode ascertainment is described in further detail in (Boulle et al., 2019; Tamuhla et al., 2021). Diabetes metrics include: 'Diabetes ascertainment age', 'Diabetes treatment initiation age' which was defined as the participant age at the recorded diabetes treatment start date, 'Diabetes treatment initiation' which was inferred from the PHDC data by calculating the time interval in years between diabetes ascertainment date and diabetes treatment start date, 'Diabetes treatment formulation' which was defined as the combination of the different classes of diabetes medication dispensed to an individual in the study observation window, 'Hypertension' which was defined as hypertension ascertainment before or during the study observation window identified by dispensing of anti-hypertensive medication, and 'Tuberculosis' which was defined as a Tuberculosis episode ascertained during the study observation window, and ‘HIV positive’ which was defined as HIV infection ascertainment before or during the study observation window.

In South Africa, HIV is managed through a parallel, vertically funded, well-resourced chronic disease programme with a large focus on adherence, which may impact adherence to other chronic disease including diabetes. Since the Khayelitsha sub-district has a high burden of HIV, to determine if there were differences in adherence based on HIV antiretroviral therapy (ART) use, we compared the sub-populations who were identified as using HIV ART vs those who were not. ‘HIV antiretroviral therapy’ use was defined as having started HIV ART before or during the study observation window.

Summary statistics were calculated for the study population. For continuous data, median and interquartile range were calculated and for grouped data, percentages were calculated.

5.3.2. Diabetes treatment persistence and adherence

The R statistical software (Team, 2020) package *AdhereR* (Dima & Dediu, 2017) was used to calculate persistence and adherence to diabetes oral drugs in the study population. ‘Persistence’ was defined as a period of continuous medication dispensing with treatment gaps of less than 90 days (Dima & Dediu, 2017; Vrijens et al., 2012). ‘Treatment gaps’ were defined as the time interval between dispensing events where no medication was dispensed, and in this study if a treatment gap was equal to or exceeded 90 days this was defined as ‘non-persistence’ or ‘treatment discontinuation’ (Dima & Dediu, 2017). If an individual was then dispensed medication following a period of ‘non-persistence’, this was treated as a new ‘treatment episode’ and the number of treatment episodes was determined by how often an individual discontinued and re-started treatment in the observation window (Dima & Dediu, 2017).

‘Adherence’ was defined as how well the treatment regimen was implemented and it was calculated for each specified medication dispensing observation window (Dima & Dediu, 2017; Vrijens et al., 2012). Since we were interested in how well the treatment regimen was implemented in the first two years following the first recorded treatment start episode, we calculated adherence to metformin in successive 4 month intervals or ‘sliding windows’ over the two year observation window (Dima & Dediu, 2017). Sliding window adherence was used to elucidate any temporal changes in adherence which would not be observed if one overall adherence measure was used (Dima & Dediu, 2017). Some patients were on more than one diabetes oral drug during the observation window, but adherence was calculated for only the drug metformin to avoid over-estimating adherence, given that *AdhereR* does not have the functionality to distinguish concurrent medication use in a treatment episode(Dima & Dediu, 2017). In this way, we used metformin as an index medication for measuring T2DM treatment adherence in PLWD.

5.3.3. Adherence trajectories

Metformin adherence trajectories for the two-year observation window were estimated from the sliding window adherence estimates using the R (Team, 2020) package *kml* (Genolini et al., 2015; Genolini & Falissard, 2010). The *kml* package applies *k*-means clustering to longitudinal data and clusters it into groups with similar characteristics (Genolini et al., 2015; Genolini & Falissard, 2010). The *k*-means clustering algorithm was implemented as previously described (Genolini et al., 2015), and the number of runs to determine each cluster partition was set to 200. Following cluster assignment, summary statistics were calculated for the sub-populations in each cluster. For continuous data, median and interquartile range were calculated and for grouped data, proportions were calculated. Wilcoxon rank sum was used to calculate significance of difference in median values between clusters, and Fisher's exact test to calculate significance of difference in proportions between clusters. Multinomial logistic regression was used to determine which factors influenced adherence trajectory in individuals in the study population, using the defined clusters as the dependent variable (outcome) and demographic and comorbidity profiles as independent variables (risk factors).

5.3.4. Monitoring treatment outcome using HbA1c

HbA1c testing is the gold standard for monitoring glycaemic control in PLWD, particularly those on treatment ("International Expert Committee Report on the Role of the A1C Assay in the Diagnosis of Diabetes," 2009; "Use of Glycated Haemoglobin (HbA1c) in the Diagnosis of Diabetes Mellitus," 2011). To determine the implementation of HbA1c testing in the study population, counts of patients with HbA1c measures were done at baseline and at six-month intervals for the duration of the two-year study observation window. Baseline HbA1c was defined as the latest HbA1c test result up to 3 months before treatment was initiated. Median HbA1c measures were also calculated for those with available data. Summary statistics were calculated for the sub-populations in each trajectory cluster. For continuous data, median and interquartile range were calculated and for grouped data, proportions were calculated. Wilcoxon rank sum was used to calculate significance of difference in median values between clusters and Fisher's exact test to calculate significance of difference in proportions between clusters.

5.3.5. Health care utilisation

The number of health facility encounters for each study participant was calculated for a three-year period starting from the six months prior to the diabetes treatment initiation until 6 months after the treatment observation window and used as a proxy for health care utilisation. During the two-year observation window, health facility encounters were counted in four month sliding windows. Counts of the total number of participants with health facility encounters were done and median encounters were also calculated.

5.3.6. Ethics

Ethics approval was granted by the University of Cape Town (HREC REF: 509/2019) and data access was approved by Western Cape Government Health (WCGH), South Africa. All data were de-identified and data perturbation was employed by the Provincial Health Data Centre (PHDC, WCGH) prior to release, so that the data used were anonymised and cannot be reidentified. Data transfer was effected through secure platforms using AES256 encryption and password protection, and analysis was undertaken on a secured, firewall-protected server. Re-use of this dataset requires approval from WCGH and contact details to apply for access are provided in the data availability statement.

5.4. Results

5.4.1. The study population

There were 16979 individuals with an inferred diabetes episode, of which 10541 met the described inclusion/exclusion criteria and were included in the study population. Summary statistics (Table 5.1) showed that the study population was 67% female with a median age for diabetes treatment initiation of 52 (IQR: 45,60) years. Most of the study population (53.7%) initiated diabetes treatment at diabetes ascertainment while 25.6% initiated treatment more than one year after diabetes ascertainment. Hypertension (61.8%), HIV (14.9%) and Tuberculosis (13.1%) were the most prevalent comorbidities with available data. In this study population, only 76.5% of people living with HIV (PLWHIV) had initiated HIV anti-retroviral treatment (Supplementary Table 5.1) before or during the two-year study observation window.

5.4.2. Diabetes treatment

PLWD in this study population were treated with metformin, sulphonylurea (gliclazide, glimepiride or glibenclamide) and insulin. During the two-year observation window, 2803 (30.2%) participants were treated with metformin only, 4109 (44.3%) with metformin and sulphonylurea and 2372 (25.5%) with metformin, sulphonylurea and insulin (Table 5.1). Comparing HIV-negative and HIV-positive groups showed no difference in the proportion of people on the different diabetes treatment formulations between the two groups in this study population (Supplementary Table 5.1). However, there was a significant difference in diabetes treatment initiation with a higher proportion of the HIV-positive group initiating treatment within one year of diabetes ascertainment compared to the HIV-negative group (Supplementary Table 5.1).

5.4.3. Diabetes treatment implementation

5.4.3.1. Persistence

An analysis of medication refill patterns in the study population in the two-year observation window showed that some individuals refilled their medication consistently while others had gaps in treatment. These gaps in treatment resulted in periods of non-persistence and multiple treatment episodes across the two years assayed which ranged in number from one to as high as five in some participants. Medication refill patterns of five study participants are illustrated in the plots in Figure 5.1 to illustrate the types of profiles seen in the study population: participant D had three treatment episodes, the first was at treatment initiation with a 28-day supply of metformin following which they had no recorded medication refill for 102 days which resulted in a period of non-persistence. Following treatment re-initiation, in the second treatment episode, the treatment formulation was changed to include a sulphonylurea (gliclazide). While this treatment episode was longer than the first, it also had medication refill gaps which resulted in the calculated adherence for the episode being 68% shown in the grey bar (Figure 5.1). The participant then had another period of non-persistence, following which they re-initiated treatment resulting in a third treatment episode. In contrast, participant B refilled their medication consistently and did not have any periods of non-persistence during the two-year observation window (Figure 5.1).

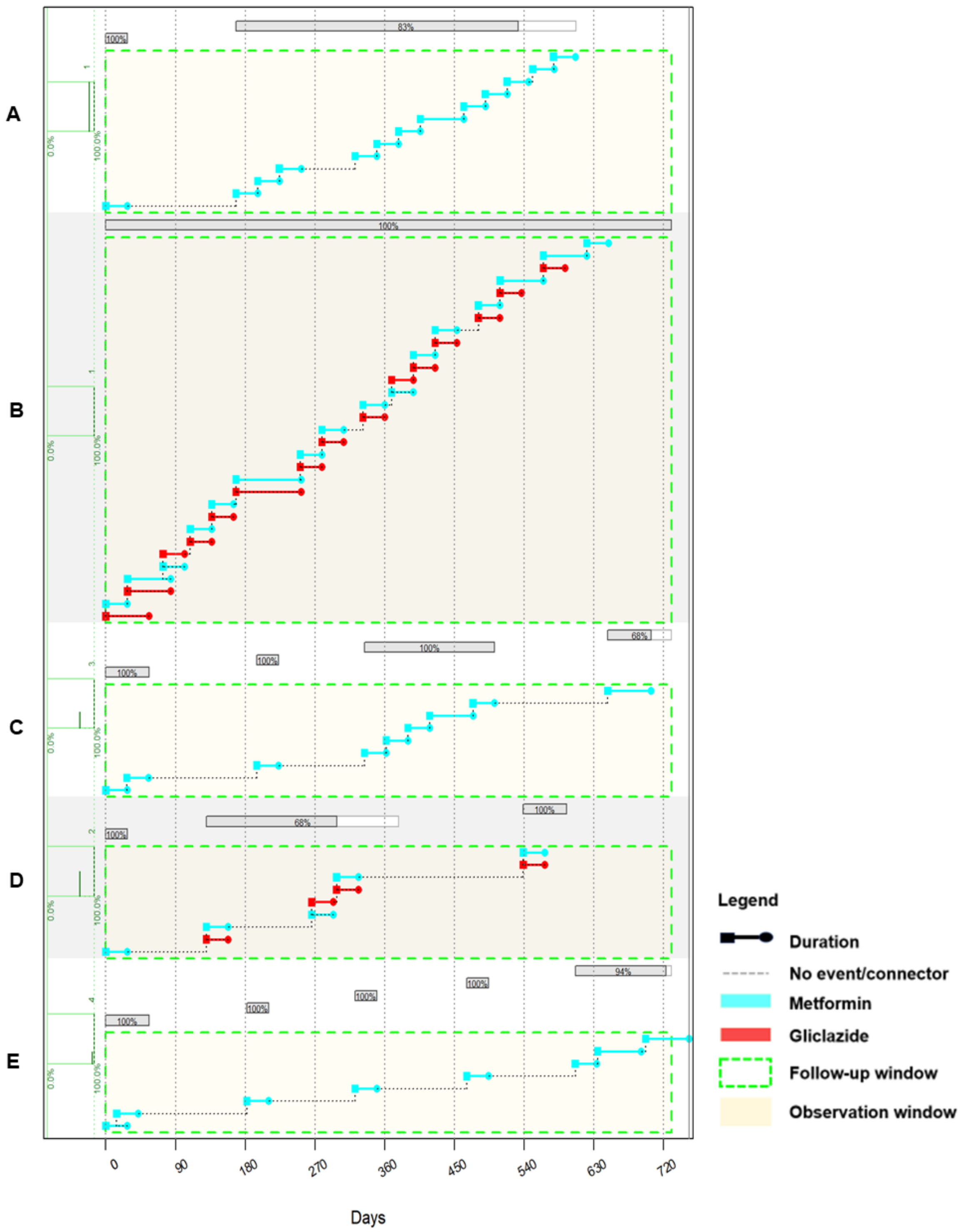


Figure 5.1. Diabetes oral medication refill patterns in five example patients in the study population. All oral diabetes medication (metformin and sulphonylurea) issued to patients during the observation window are shown.

5.4.3.2. Adherence

Since we observed that the study participants had varying medication refill patterns, we calculated adherence to metformin in the two-year observation window in 120-day sliding windows. The results in Figure 5.2 show that while some individuals, like participant B, had consistent adherence measures in the different sliding windows, others had varying levels of adherence at different time points in the two-year observation window. The 120-day sliding window adherence measures show that participant A oscillated between 17 – 35% adherence, while participant D started at 98% adherence in the first 120 days and their adherence measures gradually decreased with each subsequent sliding window to as low at 25% after 360 days on diabetes treatment (Figure 5.2).

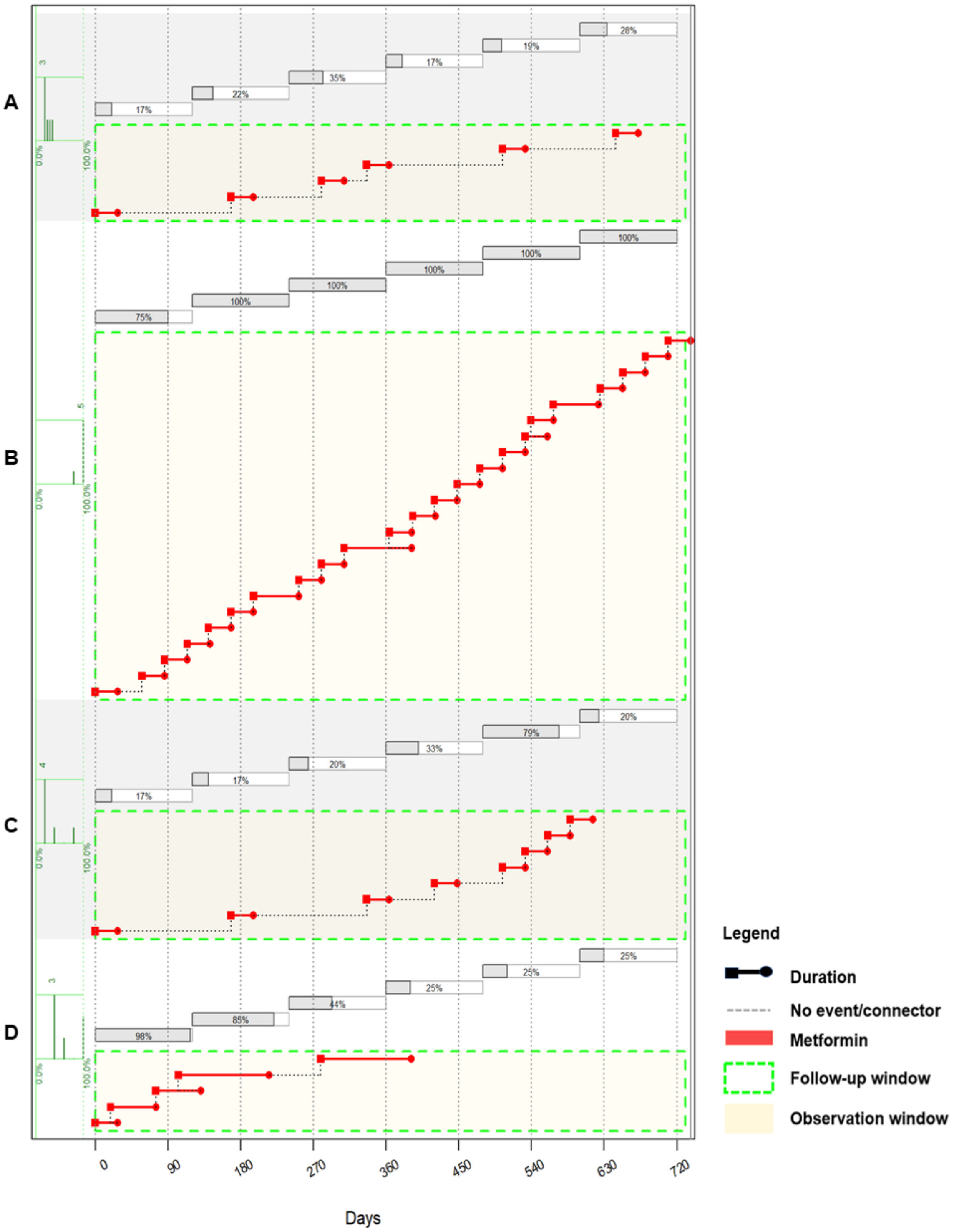


Figure 5.2. Longitudinal metformin adherence calculated in 120-day sliding windows in a two-year observation window. The results of four patients from the study population are shown.

5.4.4. Longitudinal adherence patterns

There is a general trend of decreasing adherence over time and across all sliding windows in the two-year observation window and by the end of the two-year period only 25% of the population achieved medication adherence (>= 80% adherence) (Figure 5.3A).

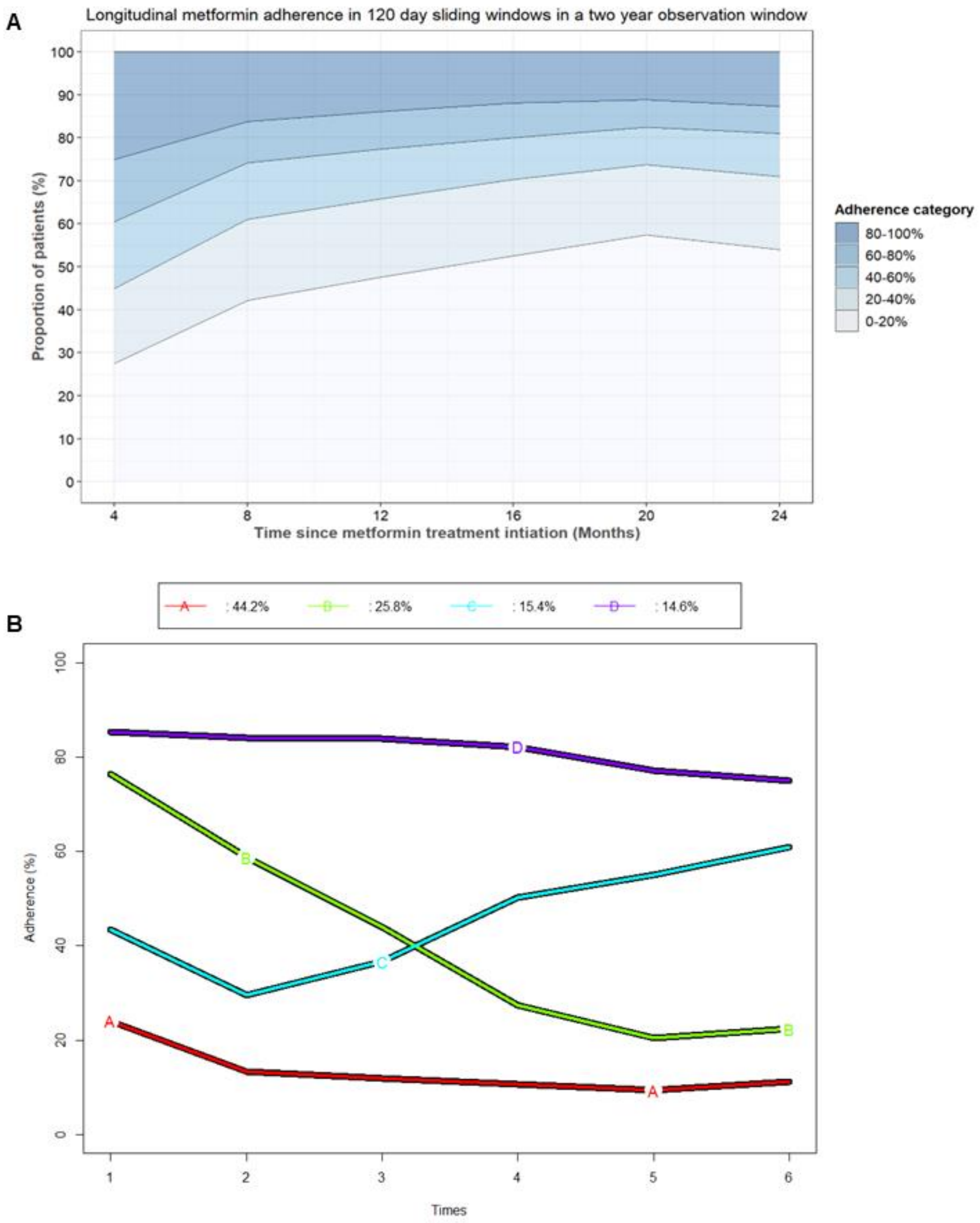


Figure 5.3. (A). A stacked area chart showing the proportion of patients in different adherence categories to visualise longitudinal trends in patient adherence. Each line represents the proportion of patients at a given time point and the colour below the line represents the corresponding adherence category. (B). Predicted two-year longitudinal

metformin adherence trajectories for individuals in the study population. Four trajectories are shown: A (Low adherence gradual decline), B (High adherence rapid decline), C (Low adherence gradual increase) and D (Adherent). Adherence is shown on the y-axis and the four-month sliding windows are represented as times 1 – 6 on the x-axis. The percentage bar shows the proportion of patients in each adherence trajectory.

In addition, the proportion of adherent patients reduced with each subsequent sliding window and the lowest numbers (11.2%) were observed at 20 months post treatment initiation. Most of the non- adherent participants had adherence measures in the 0-20% adherence category across all sliding windows over the two-year observation period and proportion of patients with 0-20% adherence measures increased steadily over time from 27.5% at 4 months to 57.4% at 20 months post treatment initiation (Figure 5.3A). In the 20-24 month sliding window, we observed a slight decrease in the proportion of non-adherent participants, however, the proportion of participants who were adherent was approximately half (12.7%) of what it was 4 months after initiating treatment (25.0%) (Figure 5.3A).

5.4.5. Longitudinal metformin adherence trajectories

The longitudinal clustering of the 120-day sliding window adherence measures for each patient in the study population produced clusters with between 2 to 6 partitions, however, there was discordance among the five-criterion used to determine cluster partition choice (Supplementary Figure 5.1). The two-cluster partition which represents the typical binary outcome of adherent or non-adherent had the highest score, but in this study, we selected the four-cluster partition as the choice that best represented the study setting (Figure 5.3B) because we are interested in a more granular classification to understand the range/levels of adherence or non-adherence. In this approach, the adherent and non-adherent categories are further broken down to create four clusters consisting of one adherent cluster and 3 that represent non-adherence at different levels. The cluster partitioning of the 4 clusters converged on one distribution, whereas the 3, 5 and 6-cluster options had alternative compositions which made them less reliable.

In the four-cluster partition, the highest proportion of study participants (44.2%) were assigned to "trajectory A" which represented individuals who initiate treatment with low adherence measure that continue to decline gradually over time. "Trajectory B" had 25.8% membership and represents participants who were adherent in the first four months of treatment (Figure 3B), but their adherence declined rapidly thereafter and continued to decline over time. "Trajectory C" which had a membership of 15.4% was made up of participants who start treatment with low adherence, but over time have a gradual improvement in adherence. "Trajectory D" which has the lowest membership (14.6%) is the only trajectory with individuals that start treatment adherent and maintain their adherence long term, even though they start to decline slightly after month 16 (Figure 5.3B). The results in Figure 3B also show that for all trajectories except D, there is a sharp decline in adherence between the 1st and 2nd sliding window, following which trajectory C has an improvement while trajectories A and B continue to decline.

5.4.6. Characteristics of the adherence trajectory sub-populations

Comparing the four adherence trajectory groups (Table 5.1) showed that there were differences in age at diabetes treatment initiation between the groups, with the adherent group (D) having the youngest median age at treatment initiation of 49 years (IQR: 42,57). There were also differences in diabetes treatment formulation between the groups with the low adherence-gradual decline group (A) having the highest proportion of individuals (43.6%) on metformin only. There were also differences in diabetes treatment initiation with the high adherence-rapid decline group (B) having the highest proportion of individuals (29.9%) who initiated treatment more than a year after diabetes ascertainment. There was also a difference in the prevalence of hypertension between the groups with the low adherence-gradual increase group (C) having the highest proportion of individuals who were ascertained with hypertension (68.1%). In addition, trajectory D, the adherent group, had the highest proportion of participants who were HIV positive (22.8%) and on HIV antiretroviral therapy (ART) (20.5%).

Table 5.1. Characteristics of the study population stratified by longitudinal metformin adherence trajectory

	Whole study population *N=10541*	**Adherent (D)** *N=1544*	**Low adherence gradual decline (A)** *N=4656*	**High adherence rapid decline (B)** *N=2716*	**Low adherence gradual increase (C)** *N=1625*
Sex: Female	7053 (67.0%)	1014 (65.8%)	3136 (67.4%)	1830 (67.5%)	1073 (66.2%)
Diabetes Ascertainment Age (Years)	52.0 [44.0;59.0]	49.0 [42.0;57.0]	53.0 [45.0;61.0]	52.0 [45.0;59.0]	51.0 [44.0;58.0]
Diabetes Treatment Initiation Age (Years)	53.0 [45.0;60.0]	50.0 [43.0;57.0]	54.0 [45.0;62.0]	53.0 [46.0;60.0]	52.0 [44.0;59.0]
Diabetes Treatment Formulation:					
Metformin only	3525 (33.4%)	274 (17.7%)	2029 (43.6%)	756 (27.8%)	466 (28.7%)

	Whole study population *N=10541*	**Adherent (D)** *N=1544*	**Low adherence gradual decline (A)** *N=4656*	**High adherence rapid decline (B)** *N=2716*	**Low adherence gradual increase (C)** *N=1625*
Metformin & Sulphonylurea	4417 (41.9%)	770 (49.9%)	1682 (36.1%)	1188 (43.7%)	777 (47.8%)
Metformin, Sulphonylurea & Insulin	2599 (24.7%)	500 (32.4%)	945 (20.3%)	772 (28.4%)	382 (23.5%)
Diabetes Treatment Initiation:					
At diabetes ascertainment	5828 (55.3%)	863 (55.9%)	2693 (57.8%)	1352 (49.8%)	920 (56.6%)
Within 1 year of ascertainment	2156 (20.5%)	313 (20.3%)	926 (19.9%)	553 (20.4%)	364 (22.4%)
More than 1 year after ascertainment	2557 (24.3%)	368 (23.8%)	1037 (22.3%)	811 (29.9%)	341 (21.0%)
HIV Antiretroviral Therapy:	1202 (11.4%)	317 (20.5%)	472 (10.1%)	236 (8.7%)	177 (10.9%)
HIV Positive:	1572 (14.9%)	352 (22.8%)	651 (14.0%)	336 (12.4%)	233 (14.3%)
Hypertension:	6517 (61.8%)	933 (60.5%)	2749 (59.0%)	1728 (63.6%)	1107 (68.1%)
Tuberculosis:	1385 (13.1%)	236 (15.3%)	647 (13.9%)	325 (12.0%)	1770.9%)

5.4.7. Predictors for longitudinal diabetes treatment adherence trajectory

Low adherence gradual decline (A) vs Adherent (D)

Treatment age (OR: 1.02, 95% CI: 1.01 – 1.03), HIV ascertainment (OR: 1.70, 95% CI: 1.16 – 2.50) or TB ascertainment (OR: 1.28, 95% CI: 1.07 – 1.53) before and/or concurrently with diabetes treatment were associated with a higher likelihood of an individual having adherence trajectory A when compared to trajectory D. HIV antiretroviral treatment (OR: 0.26, 95% CI: 0.17 – 0.39) or hypertension ascertainment (OR: 0.77, 95% CI: 0.68 – 0.87) before and/or concurrently with diabetes treatment was associated with a lower likelihood of having adherence trajectory A when compared to trajectory D. When compared to individuals on metformin only, people who were on metformin & sulphonylurea (OR: 0.29, 95% CI: 0.25– 0.34) or metformin, sulphonylurea & insulin (OR: 0.25, 95% CI: 0.21 – 0.30) were all associated with a lower likelihood of having adherence trajectory A when compared to trajectory D. Sex and time of treatment initiation in relation to ascertainment were not associated with patient adherence trajectory (Supplementary Figure 5.2).

High adherence rapid decline (B) vs Adherent (D)

Treatment age (OR: 1.01, 95% CI: 1.00 – 1.02) or HIV ascertainment before and/or concurrently with diabetes treatment (OR: 1.56, 95% CI: 1.05 – 2.31) were associated with a higher likelihood of a patient having adherence trajectory B when compared to trajectory D. When compared with initiating treatment at diabetes ascertainment, treatment initiation more than one year after diabetes ascertainment (OR: 1.43, 95% CI: 1.23 – 1.67) was associated with a higher likelihood of an individual having trajectory B when compared to trajectory D. Taking HIV antiretroviral treatment before or concurrently with diabetes treatment (OR: 0.26, 95% CI: 0.17 – 0.40) was associated with a lower likelihood of having trajectory B when compared to trajectory D. When compared to individuals on metformin only, people who were on metformin & sulphonylurea (OR: 0.54, 95% CI: 0.45 – 0.63) or metformin, sulphonylurea & insulin (OR: 0.52, 95% CI: 0.44 – 0.63) had a lower likelihood of having adherence trajectory B when compared to trajectory D. Sex, hypertension, tuberculosis and treatment initiation within one year of diabetes ascertainment were not associated with patient adherence trajectory (Supplementary Figure 5.2).

Low adherence gradual increase (C) vs Adherent (D)

Hypertension ascertainment (OR: 1.28, 95% CI: 1.10 – 1.50) before and/or concurrently with diabetes treatment was associated with a higher likelihood of a patient having adherence trajectory C when compared to trajectory D. Taking HIV antiretroviral treatment before or concurrently with diabetes treatment (OR: 0.35, 95% CI: 0.22 – 0.55) was associated with a lower likelihood of a having trajectory C when compared to trajectory D. When compared to individuals on metformin only, people who were on metformin & sulphonylurea (OR: 0.59, 95% CI: 0.49 – 0.71) or metformin, sulphonylurea & insulin (OR: 0.46, 95% CI: 0.38 – 0.57) had a lower likelihood of having adherence trajectory C when compared to trajectory D. Sex, tuberculosis, treatment initiation, treatment age and HIV were not associated with patient adherence trajectory (Supplementary Figure 5.2).

5.4.8. Monitoring glycaemic control using HbA1c

8474 study participants had at least one HbA1c measure in the two-year observation period, however, only 26.7% (2262) had a baseline HbA1c before initiating diabetes treatment (Table 2). The median HbA1c at baseline was above 9% for all adherence trajectories and the highest in trajectory D at 10.9% (IQR: 8.60, 13.0) and the lowest in trajectory A at 9.1% (IQR: 7.30, 11.3). While there was a general trend of decreasing median HbA1c across all adherence trajectories in the two-year observation, it was still above 8% across all trajectories. In addition, there was also a general trend of low HbA1c implementation with less than 35% of the study population having an available HbA1c measure in each of the sampled time periods during the two-year observation window (Table 5.2). When looking at HbA1c implementation in general, the proportion of study participants with at least one HbA1c in the first year following diabetes ascertainment was less than 90% across all adherence trajectories and there was a gradual decrease in the number of individuals with recorded HbA1c measures in each successive year and after 5 years less than 40% of the study population had a recorded HbA1c measure (Supplementary Table 5.2).

Table 5.2. HbA1c measures and proportions of study participants with available HbA1c measures in the period before treatment initiation (Baseline) and in six-month intervals during the two-year observation window.

	Whole study population *N=8474*	**Adherent (D)** *N=1442*	**Low adherence gradual decline (A)** *N=3152*	**High adherence rapid decline (B)** *N=2444*	**Low adherence gradual increase (C)** *N=1436*	**N¥**
Baseline HbA1c (%), median [IQR]	9.7 [7.6;11.9]	10.9 [8.6;13.0]	9.1 [7.3;11.3]	9.9 [7.8;12.3]	9.6 [7.8;11.7]	2262
Participants with baseline HbA1c, N (%)	2262 (26.7%)	339 (23.5%)	945 (30.0%)	585 (23.9%)	393 (27.4%)	
Six-month HbA1c (%), median [IQR]	9.1 [7.3;11.5]	9.7 [7.7;11.7]	8.6 [7.0;11.1]	9.5 [7.4;11.7]	8.8 [7.2;11.5]	2328
Participants with six-month HbA1c, N (%)	2328 (27.5%)	440 (30.5%)	769 (24.4%)	765 (31.3%)	354 (24.7%)	
One-year HbA1c (%), median [IQR]	8.7 [7.1;11.0]	8.7 [7.2;10.9]	8.4 [7.0;10.8]	8.8 [7.1;11.5]	8.9 [7.1;10.9]	2694
Participants with one-year HbA1c, N (%)	2694 (31.8%)	539 (37.4%)	887 (28.1%)	844 (34.5%)	424 (29.5%)	
Eighteen-month HbA1c (%), median [IQR]	8.5 [7.1;10.9]	8.5 [7.1;10.7]	8.6 [7.0;11.0]	8.5 [7.1;10.7]	8.6 [7.1;10.9]	2157

	Whole study population *N=8474*	**Adherent (D)** *N=1442*	**Low adherence gradual decline (A)** *N=3152*	**High adherence rapid decline (B)** *N=2444*	**Low adherence gradual increase (C)** *N=1436*	**N¥**
Participants with eighteen-month HbA1c, N (%)	2157 (25.5%)	398 (27.6%)	636 (20.2%)	632 (25.9%)	491 (34.2%)	
Two-year HbA1c (%), median [IQR]	9.0 [7.2;11.3]	9.2 [7.3;11.4]	8.9 [7.2;11.2]	8.9 [7.2;11.3]	9.1 [7.3;11.3]	2597
Participants with two-year HbA1c, N (%)	2597 (30.6%)	490 (34.0%)	859 (27.3%)	757 (31.0%)	491 (34.2%)	

¥Number of participants with available data

Across all adherence trajectories general health care utilisation was highest in the first four months after treatment initiation (Supplementary Table 5.3). Trajectory A dropped to 67.2% after 8 months and maintained a health care utilisation of between 60 -70 % for the duration of the study observation window. Trajectories B, C and D all maintained a health utilization above 80% up to 16 months in the study observation period, but by the end of the two-year period, only Trajectory C still had a health care utilisation above 80% (Supplementary Table 5.3).

5.5. Discussion

Pharmacy dispensing records from administrative health data have been widely used to estimate adherence to antidiabetic medication because they allow for the use of objective measures of adherence such as medication possession ratio (MPR) and proportion of days covered (PDC) (Chepulis et al., 2020; Kirkman et al., 2015; Lin et al., 2017; Lo-Ciganic et al., 2016; Melzer-Cohen et al., 2020; Nishimura et al., 2019). In the current study we similarly used pharmacy dispensing records as a proxy for medication use and used a modified continuous measure of medication acquisition (CMA) to estimate adherence to metformin. The modified CMA was used because it allowed for the estimation of longitudinal adherence in sliding windows (Dima & Dediu, 2017) and CMA has been shown to produce adherence estimates comparable to other commonly used measures including MPR(Hess et al., 2006). Adherence was only estimated for the drug metformin because unless contraindicated, metformin is the prescribed first line drug for the management of T2DM in South Africa (*Guidelines*, n.d.). In addition, diabetes treatment is done in an additive manner, where new drugs are added as a complement to the existing regimen (Figure 5.1), therefore estimating adherence using metformin only did not underestimate the measure in this study.

To monitor glycaemic control, the South African guidelines on the management of T2DM recommend HbA1c measures be done at the initial visit, then every 3 – 6 months annually thereafter (*Guidelines*, n.d.). However, in this study population, use of HbA1c to monitor glycaemic control was poorly implemented as most study participants had only one measure annually, while some were not having their HbA1c monitored at all (supplementary Table 5.2). This situation is not unique to this study population as similar trends of low HbA1c implementation as a monitoring tool have also been observed elsewhere in South Africa (Essel et al., 2015; Govender et al., 2017) and in Singapore (Lin et al., 2017). In the current study, we observed median HbA1c values that were increasing over time even in those who were adherent to treatment (Supplementary Table 5.2). However, it might not be that HbA1c values in the population are increasing over time, but rather this observed trend could be because those who are accessing care and getting their HbA1c tested are those who are symptomatic and have chronic hyperglycaemia (Govender et al., 2017) that is more difficult to control. In addition, because the proportion of people with recorded HbA1c was decreasing every year (Supplementary Table 5.2), this could mean that the available HbA1c values do not accurately reflect what is happening in the population since they might be biased to those who have symptomatic hyperglycaemia. Since hyperglycaemia is largely asymptomatic, there is a need for regular monitoring of glycaemic control especially in 'healthier' asymptomatic individuals to prevent early onset of diabetes-related morbidity.

Polypharmacy and comorbidities have been reported as negative predictors of adherence (Egede et al., 2014).In this study, however, we observed that individuals with chronic comorbidities that also required long term medication were likely to be adherent to their diabetes treatment. Similar findings were also observed in a study in Ethiopia where adherence to diabetes treatment increased with the number of non-diabetes medication an individual was prescribed (Rwegerera, 2014). This difference in observation maybe be because prior studies were done in older populations whereas our study population and the Ethiopian study population are generally younger with a median age of diabetes ascertainment less than 60 years (Table 5.1). In addition, non-adherence due to polypharmacy is believed to be largely linked to medication costs, and in our study population the cost of treatment is not a barrier as health care is provided for free in public health facilities in South Africa. In addition, we also observed that people who were on a complex diabetes treatment regimen were more likely to be adherent (Supplementary Figure 5.2) than those on metformin only. This may be because these are people who might have more advanced disease and are therefore symptomatic and seeking and receiving more routine care. Results from a study using patient self-report in Ethiopia, however, found that participants who were on complex regimens were less likely to be adherent as they perceived themselves to be sicker and their situation helpless (Ali et al., 2017).

Results from the current study showed that individuals with long term adherence (trajectory D) had higher median health care encounters across all sliding windows in the two-year study observation period (Supplementary Table 5.3). Similar findings were seen in other studies where individuals who had frequent access to health care were more likely to be adherent to diabetes treatment (Dobbins et al., 2019). In addition, in the current study PLWD on ART were more likely to have long term adherence. This might be because HIV infection in particular adherence to ART is managed through well-resourced programme in South Africa, therefore PLWHIV are accessing health care more (Osei-Yeboah et al., 2021) and are therefore likely to have better linkage to care for other comorbidities including diabetes. Given all these observations, it might be worth modelling the level of care given to PLWHIV in South Africa to PLWD. This could be particularly beneficial at the beginning of treatment since we observed that adherence in the first 4 months of initiating treatment was already less than 30% in this study population suggesting an urgent need for early intervention especially for those that initiate treatment while asymptomatic.

The use of routine data has made it possible to assess adherence in a very large virtual cohort of PLWD, and to understand some of the drivers of adherence across this large and diverse virtual population. Understanding real-world data in this way can provide insights into how healthcare clients access their medication and provide insights to design interventions to support healthcare clients in achieving better adherence trajectories.

5.6. References

Afroz, A., Ali, L., Karim, M. N., Alramadan, M. J., Alam, K., Magliano, D. J., & Billah, B. (2019). Glycaemic Control for People with Type 2 Diabetes Mellitus in Bangladesh—An urgent need for optimization of management plan. *Scientific Reports*, *9*(1), Article 1. https://doi.org/10.1038/s41598-019-46766-9

Aikens, J. E., & Piette, J. D. (2013). Longitudinal association between medication adherence and glycaemic control in Type 2 diabetes. *Diabetic Medicine: A Journal of the British Diabetic Association*, *30*(3), 338–344. https://doi.org/10.1111/dme.12046

Ali, M., Alemu, T., & Sada, O. (2017). Medication adherence and its associated factors among diabetic patients at Zewditu Memorial Hospital, Addis Ababa, Ethiopia. *BMC Research Notes*, *10*(1), 676. https://doi.org/10.1186/s13104-017-3025-7

Boulle, A., Heekes, A., Tiffin, N., Smith, M., Mutemaringa, T., Zinyakatira, N., Phelanyane, F., Pienaar, C., Buddiga, K., Coetzee, E., Rooyen, R. van, Dyers, R., Fredericks, N., Loff, A., Shand, L., Moodley, M., Vega, I. de, & Vallabhjee, K. (2019). Data Centre Profile: The Provincial Health Data Centre of the Western Cape Province, South Africa. *International Journal of Population Data Science*, *4*(2), Article 2. https://doi.org/10.23889/ijpds.v4i2.1143

Camara, A., Baldé, N. M., Sobngwi-Tambekou, J., André, P., Kengne, A. P., Diallo, M. M., Tchatchoua, A. P. K., Kaké, A., Ngamani, S., Balkau, B., Bonnet, F., & Sobngwi, E. (2015). Poor glycemic control in type 2 diabetes in the South of the Sahara: The issue of limited access to an HbA1c test. *Diabetes Research and Clinical Practice*, *108*(1), 187–192. https://doi.org/10.1016/j.diabres.2014.08.025

Campos, C. (2012). Chronic Hyperglycemia and Glucose Toxicity: Pathology and Clinical Sequelae. *Postgraduate Medicine*, *124*(6), 90–97. https://doi.org/10.3810/pgm.2012.11.2615

Chepulis, L., Mayo, C., Morison, B., Keenan, R., Lao, C., Paul, R., & Lawrenson, R. (2020). Metformin adherence in patients with type 2 diabetes and its association with glycated haemoglobin levels. *Journal of Primary Health Care*, *12*(4), 318–326. https://doi.org/10.1071/HC20043

Cohen, H. W., Shmukler, C., Ullman, R., Rivera, C. M., & Walker, E. A. (2010). Measurements of medication adherence in diabetic patients with poorly controlled HbA1c. *Diabetic Medicine*, *27*(2), 210–216. Scopus. https://doi.org/10.1111/j.1464-5491.2009.02898.x

Demoz, G. T., Wahdey, S., Bahrey, D., Kahsay, H., Woldu, G., Niriayo, Y. L., & Collier, A. (2020). Predictors of poor adherence to antidiabetic therapy in patients with type 2 diabetes: A cross-sectional study insight from Ethiopia. *Diabetology & Metabolic Syndrome*, *12*(1), 62. https://doi.org/10.1186/s13098-020-00567-7

Dima, A. L., & Dediu, D. (2017). Computation of adherence to medication and visualization of medication histories in R with AdhereR: Towards transparent and reproducible use of electronic healthcare data. *PLOS ONE*, *12*(4), e0174426. https://doi.org/10.1371/journal.pone.0174426

Dobbins, J. M., Elliott, S. W., Cordier, T., Haugh, G., Renda, A., Happe, L., & Turchin, A. (2019). Primary Care Provider Encounter Cadence and HbA1c Control in Older Patients With Diabetes. *American Journal of Preventive Medicine*, *57*(4), e95–e101. https://doi.org/10.1016/j.amepre.2019.04.018

Effect of intensive blood-glucose control with metformin on complications in overweight patients with type 2 diabetes (UKPDS 34). (1998). *The Lancet*, *352*(9131), 854–865. https://doi.org/10.1016/S0140-6736(98)07037-8

Egede, L. E., Gebregziabher, M., Echols, C., & Lynch, C. P. (2014). Longitudinal Effects of Medication Nonadherence on Glycemic Control. *Annals of Pharmacotherapy*, *48*(5), 562–570. https://doi.org/10.1177/1060028014526362

Erasmus, R. T., Blanco, E. B., Okesina, A. B., Gqweta, Z., & Matsha, T. (1999). Assessment of glycaemic control in stable type 2 black South African diabetics attending a peri-urban clinic. *Postgraduate Medical Journal*, *75*(888), 603–606. https://doi.org/10.1136/pgmj.75.888.603

Essel, V., van Vuuren, U., De Sa, A., Govender, S., Murie, K., Schlemmer, A., Gunst, C., Namane, M., Boulle, A., & de Vries, E. (2015). Auditing chronic disease care: Does it make a difference? *African Journal of Primary Health Care & Family Medicine*, *7*(1). https://doi.org/10.4102/phcfm.v7i1.753

Evans, M., Engberg, S., Faurby, M., Fernandes, J. D. D. R., Hudson, P., & Polonsky, W. (n.d.). Adherence to and persistence with antidiabetic medications and associations with clinical and economic outcomes in people with type 2 diabetes mellitus: A systematic literature review. *Diabetes, Obesity and Metabolism*, *n/a*(n/a). https://doi.org/10.1111/dom.14603

Fowler, M. J. (2008). Microvascular and Macrovascular Complications of Diabetes. *Clinical Diabetes*, *26*(2), 77–82. https://doi.org/10.2337/diaclin.26.2.77

García-Pérez, L.-E., Álvarez, M., Dilla, T., Gil-Guillén, V., & Orozco-Beltrán, D. (2013). Adherence to Therapies in Patients with Type 2 Diabetes. *Diabetes Therapy*, *4*(2), 175–194. https://doi.org/10.1007/s13300-013-0034-y

Genolini, C., Alacoque, X., Sentenac, M., & Arnaud, C. (2015). kml and kml3d: R Packages to Cluster Longitudinal Data. *Journal of Statistical Software*, *65*, 1–34. https://doi.org/10.18637/jss.v065.i04

Genolini, C., & Falissard, B. (2010). KmL: K-means for longitudinal data. *Computational Statistics*, *25*(2), 317–328. https://doi.org/10.1007/s00180-009-0178-4

Govender, R. D., Gathiram, P., & Panajatovic, M. (2017). Poor control and management of type 2 diabetes mellitus at an under-resourced South African Hospital: Is it a case of clinical inertia? *South African Family Practice*, *59*(5), Article 5.

Guidelines. (n.d.). Medical. Retrieved June 19, 2022, from https://www.semdsa.org.za/for-members/guidelines

Hess, L. M., Raebel, M. A., Conner, D. A., & Malone, D. C. (2006). Measurement of Adherence in Pharmacy Administrative Databases: A Proposal for Standard Definitions and Preferred Measures. *Annals of Pharmacotherapy*, *40*(7–8), 1280–1288. https://doi.org/10.1345/aph.1H018

Ho, P. M., Rumsfeld, J. S., Masoudi, F. A., McClure, D. L., Plomondon, M. E., Steiner, J. F., & Magid, D. J. (2006). Effect of medication nonadherence on hospitalization and mortality among patients with diabetes mellitus. *Archives of Internal Medicine*, *166*(17), 1836–1841. Scopus. https://doi.org/10.1001/archinte.166.17.1836

IDF Diabetes Atlas 10th Edition. (n.d.). Retrieved May 2, 2022, from https://diabetesatlas.org/data/

Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). (1998). *The Lancet*, *352*(9131), 837–853. https://doi.org/10.1016/S0140-6736(98)07019-6

International Expert Committee Report on the Role of the A1C Assay in the Diagnosis of Diabetes. (2009). *Diabetes Care*, *32*(7), 1327. https://doi.org/10.2337/dc09-9033

Khunti, N., Khunti, N., & Khunti, K. (2019). Adherence to type 2 diabetes management. *British Journal of Diabetes*, *19*(2), Article 2. https://doi.org/10.15277/bjd.2019.223

Kirkman, M. S., Rowan-Martin, M. T., Levin, R., Fonseca, V. A., Schmittdiel, J. A., Herman, W. H., & Aubert, R. E. (2015). Determinants of Adherence to Diabetes Medications: Findings From a Large Pharmacy Claims Database. *Diabetes Care*, *38*(4), 604–609. https://doi.org/10.2337/dc14-2098

Kretchy, I. A., Koduah, A., Ohene-Agyei, T., Boima, V., & Appiah, B. (2020). The Association between Diabetes-Related Distress and Medication Adherence in Adult Patients with Type 2 Diabetes Mellitus: A Cross-Sectional Study. *Journal of Diabetes Research*, *2020*, e4760624. https://doi.org/10.1155/2020/4760624

Lin, L.-K., Sun, Y., Heng, B. H., Chew, D. E. K., & Chong, P.-N. (2017). Medication adherence and glycemic control among newly diagnosed diabetes patients. *BMJ Open Diabetes Research and Care*, *5*(1), e000429. https://doi.org/10.1136/bmjdrc-2017-000429

Lo-Ciganic, W.-H., Donohue, J. M., Jones, B. L., Perera, S., Thorpe, J. M., Thorpe, C. T., Marcum, Z. A., & Gellad, W. F. (2016). Trajectories of Diabetes Medication Adherence and Hospitalization Risk: A Retrospective Cohort Study in a Large State Medicaid Program. *Journal of General Internal Medicine*, *31*(9), 1052–1060. https://doi.org/10.1007/s11606-016-3747-6

Melzer-Cohen, C., Chodick, G., Naftelberg, S., Shehadeh, N., & Karasik, A. (2020). Metabolic Control and Adherence to Therapy in Type 2 Diabetes Mellitus Patients Using IDegLira in a Real-World Setting. *Diabetes Therapy*, *11*(1), 185–196. https://doi.org/10.1007/s13300-019-00725-9

Musenge, E. M., Michelo, C., Mudenda, B., & Manankov, A. (2015, December 21). *Glycaemic Control and Associated Self-Management Behaviours in Diabetic Outpatients: A Hospital Based Observation Study in Lusaka, Zambia* [Research

Article]. Journal of Diabetes Research; Hindawi. https://doi.org/10.1155/2016/7934654

Nishimura, R., Kato, H., Kisanuki, K., Oh, A., Hiroi, S., Onishi, Y., Guelfucci, F., & Shimasaki, Y. (2019). Treatment patterns, persistence and adherence rates in patients with type 2 diabetes mellitus in Japan: A claims-based cohort study. *BMJ Open*, *9*(3), e025806. https://doi.org/10.1136/bmjopen-2018-025806

Osei-Yeboah, R., Tamuhla, T., Ngwenya, O., & Tiffin, N. (2021). Accessing HIV care may lead to earlier ascertainment of comorbidities in health care clients in Khayelitsha, Cape Town. *PLOS Global Public Health*, *1*(12), e0000031. https://doi.org/10.1371/journal.pgph.0000031

Pinchevsky, Y., Shukla, V., Butkow, N., Raal, F. J., & Chirwa, T. (2015). The achievement of glycaemic, blood pressure and LDL cholesterol targets in patients with type 2 diabetes attending a South African tertiary hospital outpatient clinic. *Journal of Endocrinology, Metabolism and Diabetes of South Africa*, *20*(2), 81–86. https://doi.org/10.1080/16089677.2015.1056468

Polonsky, W. H., & Henry, R. R. (2016). Poor medication adherence in type 2 diabetes: Recognizing the scope of the problem and its key contributors. *Patient Preference and Adherence*, *10*, 1299–1307. https://doi.org/10.2147/PPA.S106821

Rask-Madsen, C., & King, G. L. (2013). Vascular complications of diabetes: Mechanisms of injury and protective factors. *Cell Metabolism*, *17*(1), 20–33. https://doi.org/10.1016/j.cmet.2012.11.012

Raum, E., Krämer, H. U., Rüter, G., Rothenbacher, D., Rosemann, T., Szecsenyi, J., & Brenner, H. (2012). Medication non-adherence and poor glycaemic control in patients with type 2 diabetes mellitus. *Diabetes Research and Clinical Practice*, *97*(3), 377–384. https://doi.org/10.1016/j.diabres.2012.05.026

Rwegerera, G. M. (2014). Adherence to anti-diabetic drugs among patients with Type 2 diabetes mellitus at Muhimbili National Hospital, Dar es Salaam, Tanzania- A cross-sectional study. *The Pan African Medical Journal*, *17*(252), Article 252. https://doi.org/10.11604/pamj.2014.17.252.2972

Shilubane, N. H., & Cur, M. (2010). Factors contributing to poor glycaemic control in diabetic patients at Mopani District. *Curationis*, *33*(3), 43–47. https://doi.org/10.4102/curationis.v33i3.6

Tamuhla, T., Dave, J. A., Raubenheimer, P., & Tiffin, N. (2021). Diabetes in a TB and HIV-endemic South African population: Analysis of a virtual cohort using routine health data. *PLOS ONE*, *16*(5), e0251303. https://doi.org/10.1371/journal.pone.0251303

Team, R. C. (2020). *R: A Language and Environment for Statistical Computing*.

Use of glycated haemoglobin (HbA1c) in the diagnosis of diabetes mellitus. (2011). *Diabetes Research and Clinical Practice*, *93*(3), 299–309. https://doi.org/10.1016/j.diabres.2011.03.012

Vrijens, B., De Geest, S., Hughes, D. A., Przemyslaw, K., Demonceau, J., Ruppar, T., Dobbels, F., Fargher, E., Morrison, V., Lewek, P., Matyjaszczyk, M., Mshelia, C., Clyne, W., Aronson, J. K., Urquhart, J., & Team, for the A. P. (2012). A new taxonomy for describing and defining adherence to medications. *British Journal of Clinical Pharmacology*, *73*(5), 691–705. https://doi.org/10.1111/j.1365-2125.2012.04167.x

Waari, G., Mutai, J., & Gikunju, J. (2018). Medication adherence and factors associated with poor adherence among type 2 diabetes mellitus patients on follow-up at Kenyatta National Hospital, Kenya. *Pan African Medical Journal*, *29*(1), Article 1. https://doi.org/10.4314/pamj.v29i1

Zaccardi, F., Khunti, K., Marx, N., & Davies, M. J. (2020). First-line treatment for type 2 diabetes: Is it too early to abandon metformin? *The Lancet*, *396*(10264), 1705–1707. https://doi.org/10.1016/S0140-6736(20)32523-X

Zheng, Y., Ley, S. H., & Hu, F. B. (2018). Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. *Nature Reviews. Endocrinology*, *14*(2), 88–98. https://doi.org/10.1038/nrendo.2017.151

6. Chapter 6: Implementation of a genotyped virtual African population cohort: A feasibility study in the Western Cape Province, South Africa

Relevance of chapter in thesis

There is currently limited knowledge on the genetic drivers of disease in African populations as this was a previously neglected area of research. While there has been an increase in the amount of African genomic data generated in recent times, it is still an expensive undertaking and not easily scalable to large cohorts. This chapter describes a pragmatic approach to use routine health data to generate health phenotypes that can be linked to genotype data from consenting individuals. It describes a feasibility study which combines the tools and methods optimised in chapters 2 – 5 to demonstrate the implementation of a scalable and cost-effective genotyped virtual African cohort, with an approach that can be set up in under-resourced settings.

6.1. Abstract

Background

Despite recent efforts to increase the amount of African genomic data being generated, there remains a dearth of knowledge on the genetic drivers of disease in African ancestry populations which is largely driven by the prohibitive cost of undertaking large scale genomic research in Africa.

Methods

We piloted the implementation of a cost-effective scalable virtual genotyped cohort in the Western Cape Province, South Africa. Participant recruitment was done using a tiered informed consent model and we piloted DNA collection by buccal swab from consenting participants. We used micro-array genotyping to generate genotype data from buccal swab and peripheral blood DNA samples. Phenotype data was derived from the routine health data of the participants. Finally, we demonstrated the feasibility of running nested case control genome wide association studies with these data using type 2 diabetes mellitus (T2DM) and severe COVID-19 as phenotypes.

Results

We genotyped 2 267 346 single nucleotide polymorphisms (SNPs) in 459 participant samples. A total of 1782023 (78.6%) SNPs and 343 (74%) samples passed quality control (QC) and were available for further analysis. A higher proportion (80.3%) (49/61) of buccal swab samples passed QC compared to 73.8% (294/398) of blood samples. Principal component analysis showed extensive admixture in the study population. 31 known COVID-19 associated variants were identified and when comparing their occurrence in cases and controls no significant differences were observed. Similarly, 43 known T2DM variants were identified and only rs12742393 (OR: 2.49, 95% CI: 1.69 – 3.66, p-value < 0.001), rs2466293 (OR: 2.60, 95% CI: 1.72 – 3.92, p-value < 0.001) and rs9581943 (OR: 4.11, 95% CI: 2.60 – 6.51, p-value < 0.001) occurred in significantly higher counts in the T2DM cases than the controls.

Conclusion

We have described how we conceptualised and implemented a genotyped virtual population cohort in a resource constrained environment, and we are confident that this design and implementation are appropriate to scale up the cohort to a size where novel health discoveries can be made through nested case-control studies.

6.2. Introduction

Despite recent efforts to increase the amount of African genomic data being generated (Achidi et al., 2008; Choudhury et al., 2017; N. Mulder et al., 2018; N. J. Mulder et al., 2017), there is still a marked underrepresentation of African populations in genomic research (Campbell & Tishkoff, 2008; Popejoy & Fullerton, 2016). At the same time Africa is undergoing an epidemiologic shift and is experiencing an exponential increase in the prevalence of non-communicable diseases (NCDs) like T2DM (Bigna & Noubiap, 2019; Gouda et al., 2019). While it is well known that NCDs like T2DM are caused by a combination of lifestyle and genetic factors (Bertram et al., 2013; Gill et al., 2008; Hall et al., 2011; Levitt, 2008), most African studies have focused on the lifestyle drivers of theses disease (Adeniyi, Yogeswaran, Longo-Mbenza, & Goon, 2016; Adeniyi, Yogeswaran, Longo-Mbenza, Goon, et al., 2016; Amberbir et al., 2019; Manyema et al., 2015) and there remains a dearth of knowledge on their genetic drivers in African ancestry population.

While the costs of generating human genomic data have reduced significantly in recent times, in Africa, they are still a barrier to the large-scale implementation of genomic research (Adebamowo et al., 2018; *Policy Paper: A Framework for the Implementation of Genomic Medicine for Public Health in Africa | The AAS*, n.d.; Ramsay, 2012). Single nucleotide polymorphism (SNP) genotyping is a widely used cost-effective method of generating large scale genomic data however, previously available micro-array genotyping chips were not always optimal for identifying disease associated variants in African ancestry populations (Johnston et al., 2017; Popejoy & Fullerton, 2016). The recent availability of the Infinium™ H3African Consortium V2 array (H3Africa chip) (*H3Africa Chip*, n.d.) which contains novel African variants has now made it possible to generate informative genotype data for genome wide association studies (GWAS) in African genomes.

Since GWAS identify the association of genotypes with phenotypes, it is critical to ensure that phenotype definition is accurate and generated in a standardised way to avoid introducing bias which can create spurious associations ("Dissecting the Phenotype in Genome-Wide Association Studies of Psychiatric Illness," 2009; Uffelmann et al., 2021). While electronic medical records have been widely used to as a readily available cost-effective resource for generating GWAS phenotypes (Abul-Husn & Kenny, 2019; Anderson et al., 2016; Casey et al., 2016; Hoffmann et al., 2018; Kho et al., 2013; Ohno-Machado et al., 2018; Pendergrass & Crawford, 2019; Zhao et al., 2019) they are currently not widely available in African countries. However, the increasing availability of electronically captured and curated patient routine health data in African health systems (Boulle et al., 2019; Lemma et al., 2020; Wabiri et al., 2019) presents an opportunity to use these in data in African genomic studies. In addition, we have previously demonstrated how routine health data can be modelled to describe patient phenotypes at both an individual and population level (Dave et al., 2021; Tamuhla et al., 2021).

Given NCDs are predicted to surpass infections as the leading cause of morbidity and mortality in the Africa by 2030 (Bigna & Noubiap, 2019; Gouda et al., 2019) and that the genetic risk of diseases in African populations cannot be accurately predicted with the existing resources, there is an urgent need to perform large scale genomic research in Africa (Chikowore et al., 2022; Kamiza et al., 2022). The current study is intended as a proof-of-concept study demonstrating how a virtual genotyped cohort linking routine health data and genotype data is feasible for generating new research outputs and understanding disease aetiology in less-resourced environments.

6.3 Methods

6.3.1. Ethics

Ethics approval was granted by the University of Cape Town (HREC REF: 509/2019) and permission to conduct the study at Groote Schuur Hospital diabetes clinics and access participant routine health data was granted by the Western Cape Government Health (WCGH), South Africa. A tiered informed consent model was used (Nembaware et al., 2019; Tamuhla et al., 2022) and the participant information and consent forms were in provided in both English (Supplementary file 6.1) and isiXhosa (Supplementary file 6.2) to ensure that language was not a barrier for prospective participants. Genomic data was de-identified using bar-coding of consent forms and collection tubes prior to recruitment and sample collection. Using the barcoded consent forms with the recorded Clinicom folder number, the Provincial Health Data Centre (PHDC, WCGH) (Boulle et al., 2019) then linked clinical records to the barcode number using the clinical folder number and returned de-identified data with only the barcode as an identifier. This facilitated linkage of data without overt exposure of personal participant details adds an additional layer of privacy protection, even though permission was provided by participants for the use of their identified data. Data transfer was effected through secure platforms using AES256 encryption and password protection, and analysis was undertaken on a secured, firewall-protected server.

6.3.2 Study population and sampling

All adults (18 years or older at recruitment) who consented to access of their clinical data from the PHDC and to give a DNA sample were eligible for participation in the study. DNA samples were collected using buccal swabs and two swabs were collected from each consenting participant. The buccal swab method was chosen because it is relatively inexpensive, easy to administer, non-invasive and the samples are easy to store and transport after collection (Matimba et al., 2008). DNA samples for genotyping were prepared from the buccal swabs at an external commercial facility (Central Analytic Facility, Stellenbosch University) and this approach was chosen to allow for the optimisation of DNA preparation protocols that can be scaled for large sample sizes. DNA quantification was done using the Qubit dsDNA High Sensitivity (HS) Assay kit according to the manufacturer's instructions.

In addition, since we were piloting a study design that would allow the cohort to grow over time, we piloted a collaborative recruitment strategy where different groups with consenting participants can collaborate and combine their DNA samples for genotyping in the same batch to increase sample size and reduce the cost. To test this approach, DNA extracted from peripheral blood samples of consenting participants from the HIATUS study was included for genotyping in this study. Participant recruitment into the HIATUS study has been described in detail elsewhere (du Bruyn et al., 2023).

6.3.3. Genotyping

DNA samples from 459 unrelated study participants were genotyped on the Infinium™ H3Africa Consortium Array V2 (H3Africa chip), a custom genotyping chip with ~2.26 million SNPs including novel African variants (*H3Africa Chip*, n.d.). Array BeadChips were analysed on the Illumina iScan™ System and the GenomeStudio™2.0 Genotyping Module was used to make genotype calls and generate PLINK PED and MAP files (Purcell et al., 2007) from the raw genotype data.

6.3.4. Genotyping quality control

Quality control (QC) of the genotyped data was done in PLINK1.9 (Purcell et al., 2007) using the protocol from Marees and colleagues (Marees et al., 2018) with some modifications. QC was done for both samples (n = 459) and SNPs (n = 2 267 346).

Sample QC

Samples with a genotype call rate of less than 98% were excluded from the dataset. Discordant sex information is when the recorded sex and the genotype sex do not match and for this analysis sex imputation using the genotype sex was done for samples with discordant sex information (Marees et al., 2018). All samples failing sex imputation were excluded from the dataset. Samples were checked for relatedness and those with an identity by descent (IBD) score of more than 0.2 were also excluded (Marees et al., 2018). Finally, samples with an outlying heterozygosity score (more than 3 standard deviations from the mean) were also excluded from the dataset (Marees et al., 2018).

SNP QC

SNPs with a missingness of more than 2%, a Hardy-Weinberg equilibrium p-value less than 1 × 10-6 and a minor allele frequency (MAF) less than 1% were all excluded from the dataset. A MAF threshold of 1% was used instead of the widely used 5% because the H3Africa chip contains some variants that occur with a MAF =< 1% in African populations (*H3Africa Chip*, n.d.) and the same threshold has been applied in a recent study (Choudhury et al., 2022) which also used genotype data from the H3Africa chip. In addition, all non-autosomal SNPS (X, Y and mitochondrial) were also excluded from the data during QC (Marees et al., 2018).

6.3.5. Population stratification

The structure of the study population was assessed using multi-dimensional scaling (MDS) and principal component (PC) analysis in PLINK1.9 (Purcell et al., 2007) following the protocol from Marees and colleagues (Marees et al., 2018). Prior to conducting the MDS and PC analysis, the data were first LD pruned leaving only a subset of uncorrelated SNPs which were then merged with the 1000 Genomes data which contains well defined reference populations. Following the analysis, 10 MDS components were extracted as covariates to be used in the genome wide association analysis to control for population stratification bias. MDS and PCA plots were generated using R statistical software (Team, 2020).

6.3.6. Imputation

Prior to phasing and imputation, reference allele mis-match was checked and any problematic data subsequently fixed and normalised using BCFtools (Danecek et al., 2021). Phasing using EAGLE2 (Loh et al., 2016) and genome wide imputation using positional Wheeler-Burrows transform (PWBT) (Durbin, 2014) were then done on the Sanger Imputation Service using the African Genome Resources reference panel (*Sanger Imputation Service - Wellcome Sanger Institute*, n.d.).

6.3.7. Nested case-control GWAS

6.3.7.1. Identification of cases and controls

To demonstrate the feasibility of carrying out a GWAS with these data, two nested case-control GWAS (T2DM and severe COVID-19) were done using the pre-imputation quality-controlled data. This was undertaken as a proof-of-principle analysis recognising that the studies would not have sufficient statistical power to generate decisive results. The cases and controls for the two studies were identified using phenotype data inferred from the PHDC records of the study population. A T2DM case was inferred from PHDC records using listed disease evidences of at least one glycated haemoglobin (HbA1c) value greater than or equal to 6.5% ("Use of Glycated Haemoglobin (HbA1c) in the Diagnosis of Diabetes Mellitus," 2011) and/or dispensed diabetes drugs as previously described (Tamuhla et al., 2021). A severe COVID-19 case was also inferred from PHDC records using listed disease evidences of a positive SARS-CoV-2 polymerase chain reaction (PCR) laboratory result and hospital admission for the treatment of SARS-CoV-2 infection as previously described (Dave et al., 2021). The participants who did not meet the defined case criteria were then treated as controls.

6.3.7.2 GWAS

Since the phenotypes in both studies were binary, A logistic regression GWAS using the 10 MDS components as covariates was done in PLINK1.9 (Purcell et al., 2007). Quantile-quantile (QQ) plots were plotted in R to check for biases in the data which if not controlled for could result in erroneous false positive associations (Marees et al., 2018). Manhattan plots were also plotted in R to identify SNPs with the strongest associations based on -log10 of their p-values.

6.3.8 Identification of known T2DM and COVID-19 variants

To do a descriptive analysis in this study population for known disease-associated variants previously identified in other populations, known T2DM and COVID-19 SNPs were identified from the literature, and compiled into lists. Using PLINK1.9 (Purcell et al., 2007) SNPS from the list were identified in the study dataset and extracted from the genotyped data, and their allelic counts and associated odds ratios calculated to determine their occurrence in the cases compared to the controls for each disease phenotype.

6.3.9. Statistical analysis

Summary statistics were calculated for the study population using R version 3.6.3 (Team, 2020). For continuous data, median and interquartile range were calculated and for grouped data, percentages were calculated. For median values, the Wilcoxon rank sum test was used to calculate significance of differences between groups; and significance of the differences in proportions between groups was tested using the Fisher's exact test. The Bonferroni correction was applied to adjust for multiple testing.

6.4. Results

6.4.1. DNA quality from buccal swabs

We collected two buccal swabs samples from 61 consenting participants and the DNA was extracted and stored separately for each sample. DNA from 49 (80.3%) participants was successfully extracted from both buccal swabs while in 12 (19.7%) participants only 1 of the buccal swabs gave an adequate DNA sample. Qubit quantification of the extracted DNA gave a median concentration of 36.9 (IQR: 23.0, 57.5) ng/µl and total median DNA yield of 2.79 (IQR: 1.88, 4.27) µg per sample. In addition, the highest DNA yields were obtained from patient self-administration under clinician supervision (data not shown).

6.4.2. Genotyping quality control

The total genotype call rate for the study was 97.0% before QC. A total of 1782023 (78.6%) SNPs and 343 (74%) samples with a final total genotype call rate of 99.9% passed QC and were available for further analysis. When comparing the QC pass rate by sample type, a higher proportion of buccal swab samples (80.3%) (49/61) passed QC compared to 73.8% (294/398) of blood samples (Table 6.1) and most of the excluded samples (n= 95) were due to excessive genotype missingness at a 2% threshold. For individual SNPs, 12.6% (n= 285975) of the genotyped SNPs were excluded because they had a MAF less than the 1% threshold.

6.4.3 Population structure

An analysis of the population structure using principal components (Figure 6.1) and multi-dimensional scaling (MDS) (Supplementary Figure 6.1) showed extensive admixture in the study population. This is because while some of the study population clustered with the African population in the 1000 Genomes data, a significant proportion did not form clusters and were spread across the plot, showing extensive genetic variation.

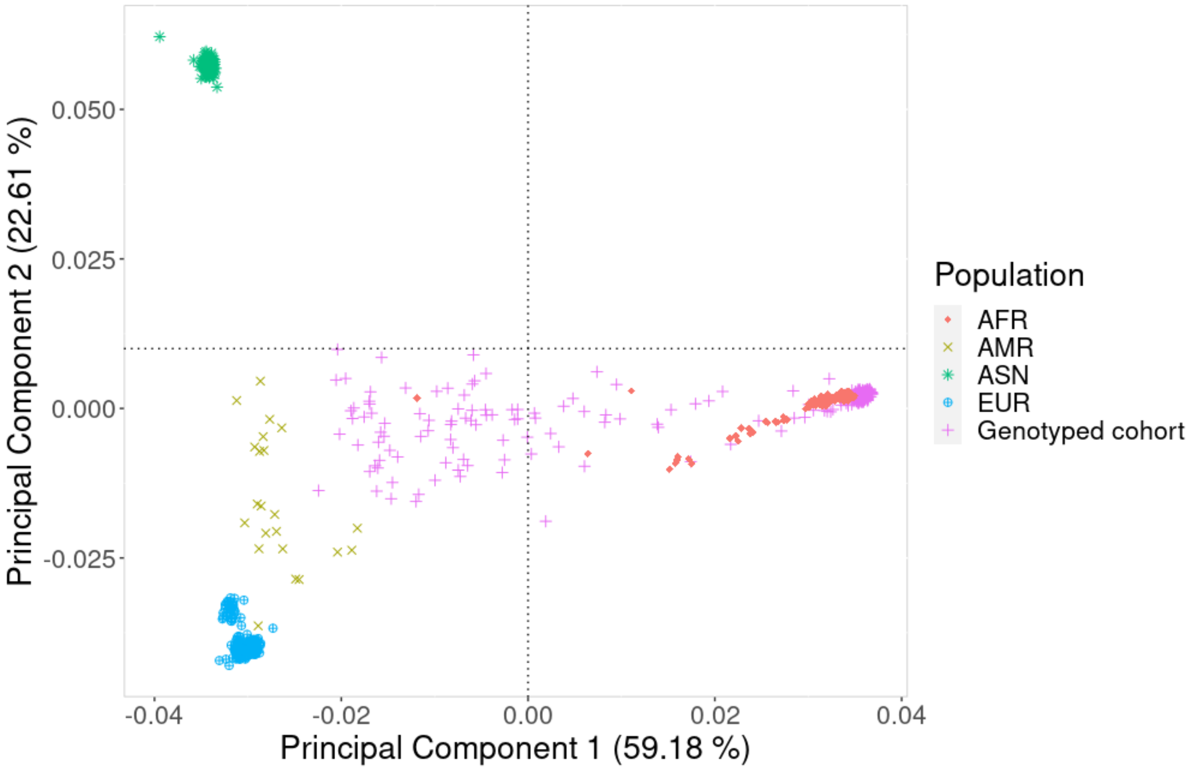


Figure 6.1. Principal component (PC) analysis plot using PC1 and PC2. These two components show the genetic variation of the study population in comparison to well defined continental populations in the 1000 Genomes dataset. AFR (♦) is African, AMR (x) is American, ASN (*) is Asian, EUR (⊕) is European and Genotyped cohort (♦) is the study participants from the Western Cape Province, South Africa.

6.4.4. Study population in the nested case-control GWAS

The 343 people who passed QC were predominantly female 66.8% (n = 229) and had a median age of 45 years [IQR: 35, 56] (Table 6.1).

When looking at each case-study separately, there were 63 cases of severe COVID-19 identified and 280 controls (Table 6.1). When comparing cases and controls there were significant differences (p< 0.001) in sex distribution and the controls had a high proportion of females (71.4%) whereas the cases were mostly male (54.0%) (Table 6.1). Similarly, there were significant differences (p<0.001) in median age between cases and controls. The cases had a median age of 54.0 years [IQR: 44.0, 64.5] compared to a younger control population of 43.0 years [IQR: 34.0, 54.0] (Table 6.1). There was no significant difference (p=0.842) between the cases and controls when looking at the DNA sample types that generated the genotype data in the severe COVID-19 GWAS (Table 6.1).

Table 6.1. Characteristics of the study participants in the severe COVID-19 nested case-control GWAS stratified by sample collection (buccal swab and peripheral blood)

	All (n = 343)	**Controls (n = 280)**	**Cases (n = 63)**	**p-value**
Sex:				<0.001
Male	114 (33.2%)	80 (28.6%)	34 (54.0%)	
Female	229 (66.8%)	200 (71.4%)	29 (46.0%)	
Age (years)	45.0 [35.0;56.0]	43.0 [34.0;54.0]	54.0 [44.0;64.5]	<0.001
DNA sample:				0.842
Blood	294 (85.7%)	241 (86.1%)	53 (84.1%)	
Buccal swab	49 (14.3%)	39 (13.9%)	10 (15.9%)	

For the T2DM GWAS, 93 cases and 250 controls were identified from the 343 samples that passed QC (Table 6.2). While the controls had a higher proportion of females (69.2%) compared to the cases (60.2%), there were no significant differences (p=0.149) in the sex distribution between the cases and controls (Table 6.2). There were however significant differences in age (p-value < 0.001) with the controls being younger with a median age of 40.5 years [IQR: 32.0, 50.0] while the median age for the cases was 57.0 years [IQR: 49.0, 65.0]. There were also significant differences in the DNA samples that generated the genotype data for the cases and controls. All the controls were from blood samples whereas the cases were genotyped from almost equal proportions of blood and buccal swab samples (Table 6.2). Since the genotyping was done in the same batch, there is no need to control for the DNA sample type.

Table 6.2. Characteristics of the participants in the T2DM nested case-control GWAS stratified by sample collection (buccal swab and peripheral blood).

	All (n = 343)	**Controls (n = 250)**	**Cases (n = 93)**	**p-value**
Sex:				0.149
Male	114 (33.2%)	77 (30.8%)	37 (39.8%)	
Female	229 (66.8%)	173 (69.2%)	56 (60.2%)	
Age (years)	45.0 [35.0;56.0]	40.5 [32.0;50.0]	57.0 [49.0;65.0]	<0.001
DNA sample:				<0.001
Blood	294 (85.7%)	250 (100.0%)	44 (47.3%)	
Buccal swab	49 (14.3%)	0 (0.0%)	49 (52.7%)	

6.4.5 Severe COVID-19 GWAS

Since the phenotype was binary, a logistic regression GWAS was done to identify variants associated with severe COVID-19 in our study population. From the results in Figure 6.2, no genotyped variants in the pre-imputation dataset were associated with severe COVID-19. This is because none had a p-value that crossed the genome wide significance threshold of 5x10^{-8}. However, 11 SNPs crossed the suggestive threshold p-value of 1x10^{-5} (Figure 6.2). In addition, the annotated SNPs on the Manhattan plot had p-values smaller than the suggestive threshold, and when multiple SNPs on the same chromosome cross the threshold only the one with the smallest p-value is annotated (Figure 6.2).

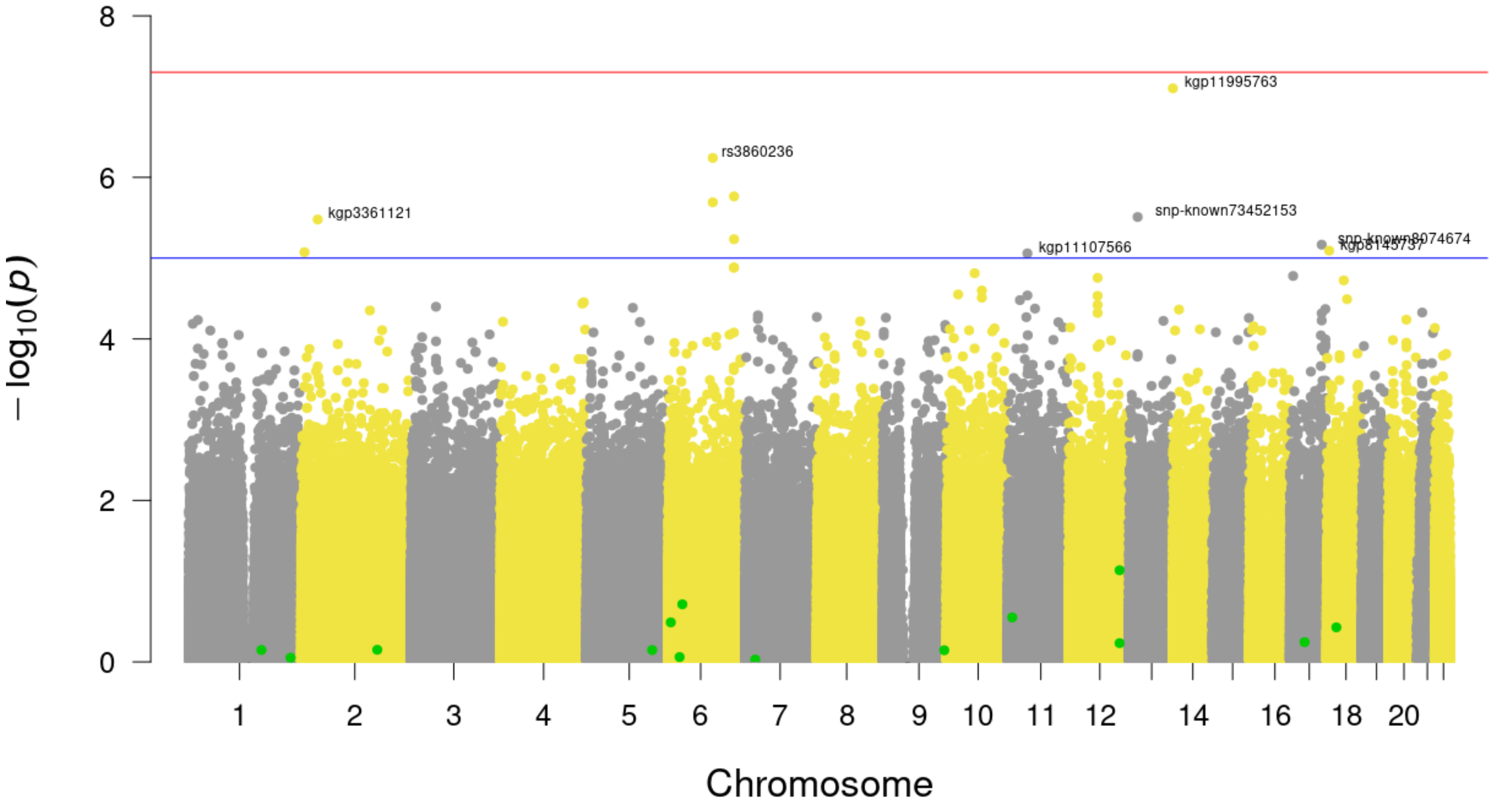


Figure 6.2. Manhattan plot of a severe COVID-19 case-control GWAS in 343 individuals from a virtual genotyped cohort in the Western Cape Province, South Africa. The genome wide significant threshold ($5x10^{-8}$) is shown by the top horizontal line and the suggestive threshold ($1x10^{-5}$) by the bottom horizontal line. The following annotated SNPs; kgp3361121, rs38660236, kgp11107566, kgp11995763, snp-known73452153, snp-known8074674 and kgp8145737 crossed the suggestive threshold. The SNPs highlighted in green represent known COVID-19 variants that were identified in the pre-imputation quality control dataset.

A quantile-quantile (QQ) plot (Figure 6.3) of the -$\log_{10}$ p-values from the Manhattan plot in Figure 6.2 showed that the observed associations were not confounded by the population admixture (Figure 6.1) that is present in the study population.

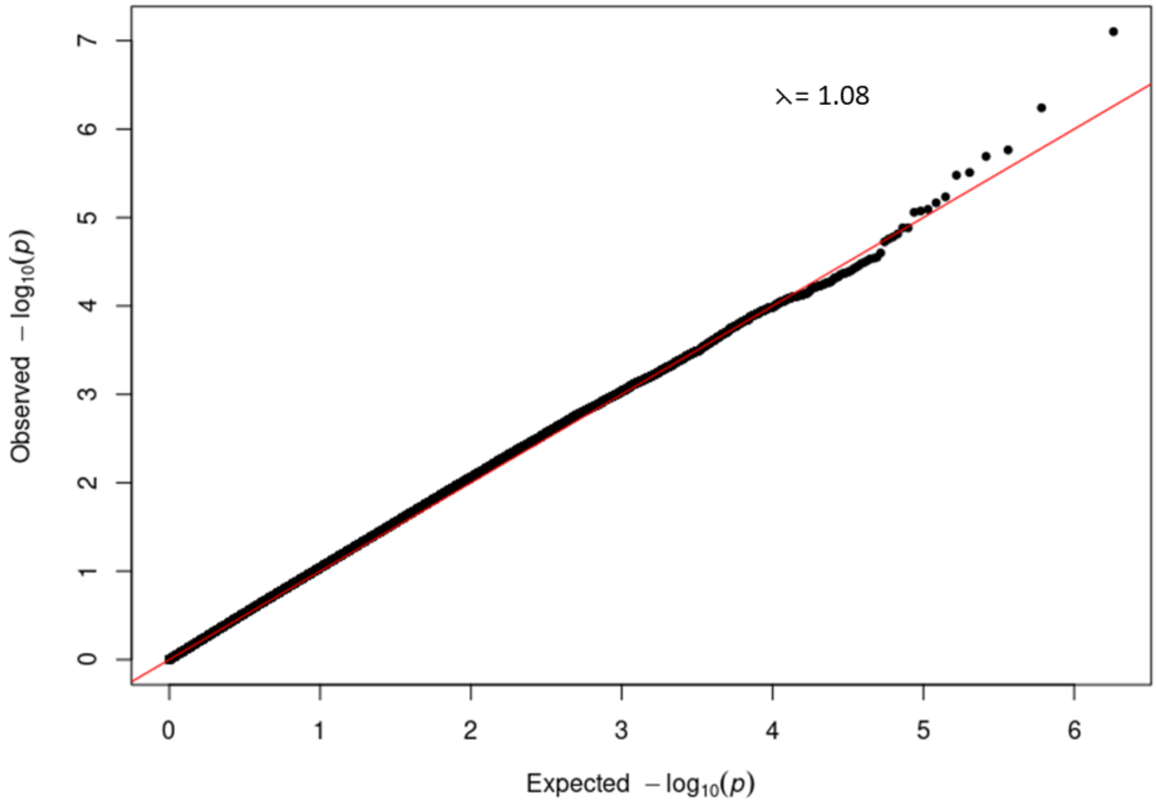


Figure 6.3. Quantile-quantile (QQ) plot of -log10 p-values from the severe COVID-19 GWAS Manhattan plot

6.4.6 T2DM GWAS

Since the phenotype was binary, a logistic regression GWAS was done to identify variants associated with T2DM in our study population. From the results in Figure 6.4, no genotyped variants in the pre-imputation dataset were associated with T2DM. This is because none had a p-value that crossed the genome wide significance threshold of 5x10^{-8}. However, 4 SNPs crossed the suggestive threshold p-value of 1x10^{-5} and are annotated on the plot (Figure 6.4).

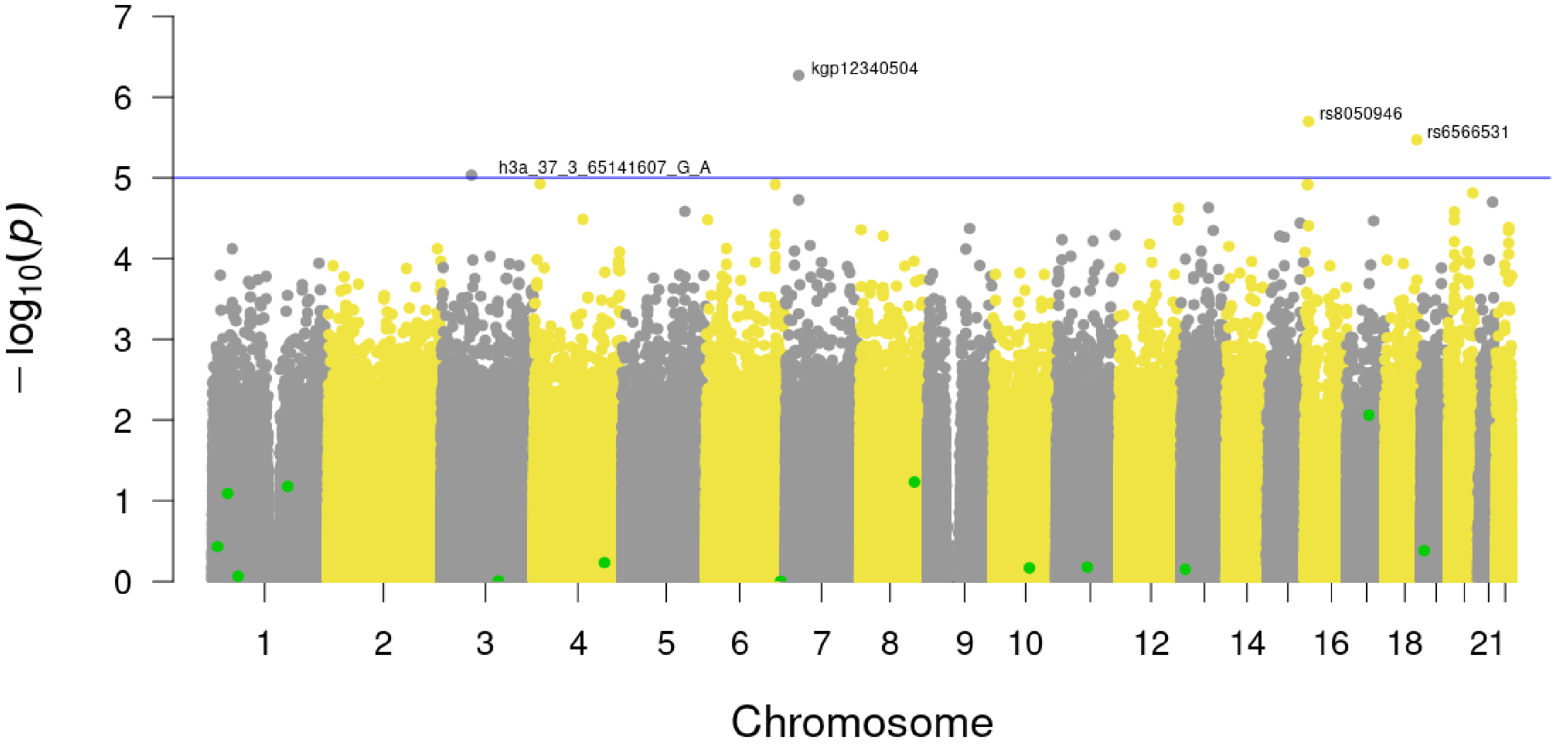


Figure 6.4. Manhattan plot of the T2DM case-control GWAS in 343 individuals from a virtual genotyped cohort in the Western Cape Province, South Africa. The suggestive threshold (1x10^{-5}) is shown by the horizontal line and the following annotated SNPs; h3a_37_65141607_GA, kgp12340504, rs8050946 and rs6566531crossed it. The SNPs highlighted in green represent known T2DM variants that were identified in the pre-imputation quality control dataset.

A quantile-quantile (QQ) plot (Figure 6.5) of the -$\log_{10}$ p-values from the Manhattan plot in Figure 6.4 showed that the observed associations were not confounded by the population admixture (Figure 6.1) that is present in the study population.

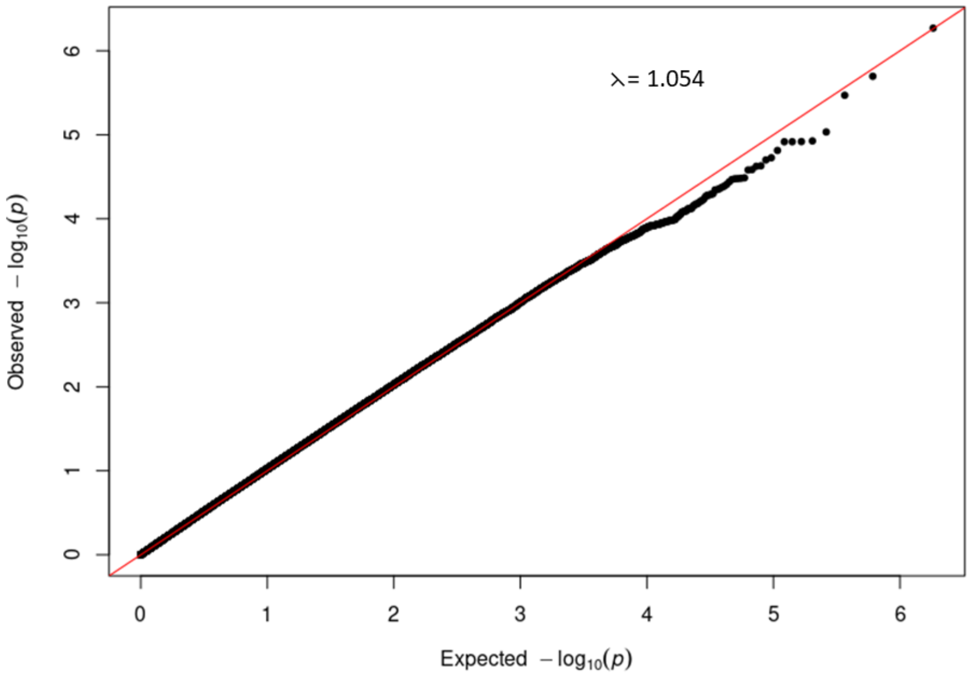


Figure 6.5. Quantile-quantile (QQ) plot of -$\log_{10}$ p-values from the T2DM GWAS Manhattan plot

6.4.7. Identification of known COVID-19 associated variants

A total of 31 known COVID-19 variants were identified in the pre-imputation dataset and of these, 17 SNPs had an allele count of 0 in both cases and controls (Supplementary Table 6.1) meaning that they were not present in any of the samples in our study. A comparison of the occurrence of the remaining 14 COVID-19 associated variants in cases and controls identified 6 SNPs associated with reduced odds of occurring in severe COVID-19 cases (odds ratio < 1), however, after adjusting for multiple-testing these odds were not statistically significant (Table 6.3). Similarly, the 8 SNPs associated with increased odds of occurring in severe COVID-19 cases did not have statistically significant odds after adjusting for multiple testing. These 14 COVID-19 associated variants occurring in our study population were highlighted in green on the GWAS Manhattan plot (Figure 6.1) and after the multiple testing adjustment in this genome wide hypothesis testing, they are not highlighted as potential aetiological variants. A larger sample size would be required to make reliable conclusions in this scenario.

Table 6.3. Allele counts of known COVID-19 associated variants identified in the feasibility study pre-imputation data.

SNP	Nearest gene	Base pair location[a]	Risk allele[b]	Risk allele counts Cases (n= 63)	Risk allele counts Controls (n = 280)	Odds ratio	95% CI[c]	p-value unadjusted[d]	p-value adjusted[e]
rs1801274	FCGRA2	Chr 1, 161479745	A	58	282	0.84	0.57-1.24	0.380	1
rs360102	TMEM63A	Chr 1, 226067862	A	46	168	1.33	0.89-2.00	0.162	1
rs7595310	STK39	Chr 2, 168810137	G	50	153	1.75	1.17-2.62	0.006	0.084
rs17448496	PPP2RB2	Chr 5, 146015615	G	16	44	1.71	0.93-3.13	0.082	1
rs17142392	LY86	Chr 6, 6626983	C	48	217	0.97	0.65-1.45	0.892	1
rs1799945	HFE	Chr 6, 26091179	G	6	10	2.75	0.98-7.71	0.046	0.637
rs3131294	NOTCH4	Chr 6, 32180146	A	1	7	0.63	0.08-5.18	0.666	1
rs2069837	IL6	Chr 7, 22768027	G	19	80	1.06	0.62-1.83	0.818	1
rs657152	ABO	Chr 9, 136139265	C	59	248	1.11	0.75-1.63	0.604	1
rs2923084	CAND1.11	Chr 11, 10388782	G	51	234	0.95	0.64-1.40	0.788	1
rs10774671	OAS1	Chr 12, 113357193	G	54	289	0.70	0.48-1.04	0.076	1
rs10735079	OAS3	Chr 12, 113380008	G	40	128	1.55	1.02-2.38	0.040	0.562
rs1024611	MCP-1	Chr 17, 32579788	G	28	86	1.57	0.98-2.54	0.061	0.860
rs4800182	OSBPL1A	Chr 18, 21812972	G	39	209	0.75	0.49-1.13	0.170	1

a. Location of the SNP on the genome denoted by chromosome number and base pair position
b. Alleles aligned to genome build GrCh37
c. 95% confidence interval
d. Fisher's exact test p-value
e. Bonferroni corrected p-value

6.4.8. Identification of known T2DM variants

A total of 43 known T2DM variants were identified in the pre-imputation dataset and of these, 30 SNPs had an allele count of 0 in both cases and controls (Supplementary Table 6.2) meaning that they were not present in any of the samples in our study. A comparison of the occurrence of the remaining 13 T2DM associated variants in cases and controls identified 2 SNPs associated with reduced odds of occurring in T2DM cases (odds ratio < 1) and after correcting for multiple testing, neither occurred in significantly lower counts in T2DM cases than controls. The other 11 SNPs were associated with increased odds of occurring in T2DM cases and of these, only rs1801133 (OR: 2.26, 95% CI: 1.29 – 3.97, p-value = 0.048), rs2495477 (OR: 1.71, 95% CI: 1.21 - 2.42, p-value = 0.028), rs12742393 (OR: 2.49, 95% CI: 1.69 – 3.66, p-value < 0.001), rs10033601 (OR: 1.66, 95% CI: 1.18 – 2.32 , p-value = 0.043), rs2466293 (OR: 2.60, 95% CI: 1.72 – 3.92 , p-value < 0.001) and rs9581943 (OR: 4.11, 95% CI: 2.60 – 6.51, p-value < 0.001) occurred in significantly higher counts in the T2DM cases than controls in our study population. In addition, the 13 T2DM associated variants occurring in our study population are highlighted on the GWAS Manhattan plot (Figure 6.4) but due to the multiple testing adjustment in this genome wide hypothesis testing, they are not highlighted as potential aetiological variants. A larger sample size would be required to make reliable conclusions in this scenario.

Table 6.4. Allele counts of known T2DM associated variants identified in the genotyped cohort pre-imputation data.

SNP	Nearest gene	Base pair location[a]	Risk[b] Allele	Risk allele counts Cases (n= 93)	Risk allele counts Controls (n = 250)	Odds Ratio	95% CI[c]	p-value unadjusted[d]	p-value adjusted[e]
rs1801133	MTHFR	Chr 1, 11856378	A	24	31	2.26	1.29-3.97	0.004	0.048
rs3480	FNDC5	Chr 1, 333328165	A	100	235	1.31	0.93-1.84	0.115	1
rs2495477	PCSK9	Chr 1, 55518467	A	83	160	1.71	1.21-2.42	0.002	0.028
rs12742393	NOS1AP	Chr 1, 162224586	C	61	82	2.49	1.69-3.66	2.61 x 10^{-6}	3.39 x 10^{-5}
rs11708067	ADCY5	Chr 3, 123065778	G	24	66	0.97	0.59-1.61	0.918	1
rs10033601	FBXW7	Chr 4, 153252061	A	103	214	1.66	1.18-2.32	0.003	0.043
rs622342	SLC22A1	Chr 6, 160572866	C	44	97	1.29	0.85-1.93	0.220	1
rs2466293	SLC30A8	Chr 8, 118185938	G	52	65	2.60	1.72-3.92	3.65 x 10^{-6}	4.75 x 10^{-5}
rs2637248	LRMDA	Chr 10, 78273721	A	57	152	1.01	0.70-1.46	0.950	1
rs1695	GSTP1	Chr 11, 67352689	G	76	263	0.62	0.44-0.88	0.006	0.081
rs9581943	PDX1	Chr 13, 28493996	A	49	40	4.11	2.60-6.51	2.06 x 10^{-10}	2.68 x 10^{-9}
rs16948048	ZNF652	Chr 17, 47440466	G	51	130	1.07	0.74-1.57	0.708	1
rs1799817	INSR	Chr 19, 7125297	A	51	108	1.37	0.93-2.02	0.108	1

a. Location of the SNP on the genome denoted by chromosome number and base pair position
b. Alleles aligned to genome build GrCh37
c. 95% confidence interval
d. Fisher's exact test p-value
e. Bonferroni adjusted p-value

6.5. Discussion

In this analysis we conducted a feasibility study demonstrating the implementation of a scalable and cost-effective virtual genotyped population cohort suitable for doing genomic research in under resourced settings. While similar cohorts have been set up elsewhere (Brumpton et al., 2022; Forgetta et al., 2022; Hewitt et al., 2016; Matimba et al., 2008; Nagai et al., 2017), the proposed virtual cohort differs in that it does not require complex infrastructure for biobanking large collections of study samples. This is because the tiered informed consent model used provides an option to recontact participants for future studies where more complex samples and data might be needed. Therefore, a participant only needs to give a DNA sample once and does not need to return for follow-up visits thus working to also reduce participant research fatigue.

We demonstrated that the virtual cohort design is an inclusive model which can incorporate collaborators from different research environments if appropriate informed consent is in place. We successfully piloted genotyping samples from different studies in the same batch and showed that samples from different sources and protocols can be combined and their routine health data used for immediate meta-analyses without having to go through a harmonisation process (Table 6.1 and 6.2).

In this feasibility study the SNP and sample QC thresholds were adopted from a protocol using European ancestry data (Marees et al., 2018) and we acknowledge that the thresholds used may not have been optimal for data generated from the H3Africa chip (N. Mulder, 2017). In particular instead of the proposed 5% MAF threshold (Marees et al., 2018) for SNP QC, we modified the protocol and used a MAF threshold of 1% in line with other studies conducting genotyping QC on African samples (Choudhury et al., 2022; May et al., 2013). As more samples get genotyped on the H3Africa chip we expect to be able to optimise suitable QC thresholds that will provide high quality data for African GWAS.

It is well established that African ancestry populations including those in our genotyped cohort (Figure 6.1) are genetically diverse (Choudhury et al., 2018, 2021; Kamiza et al., 2022; Petersen et al., 2013). While we had initially endeavoured to only include self-identifying isiXhosa speaking individuals to keep the study population homogenous, the study was set in Western Cape Province which has a highly heterogenous population (Choudhury et al., 2021). This heterogeneity was observed in the population structure analysis which showed significant stratification and admixture (Figure 6.1). We were able to demonstrate (Figure 6.5 and 6.5) that genetic and genomic analyses done with this cohort can, with the appropriate analysis tools, accommodate the enormous genomic variety in the population of the Western Cape, which has ancient and modern non-admixed and highly admixed African populations, as well as admixture from Europe and Asia (Chimusa et al., 2013; Choudhury et al., 2021; de Wit et al., 2010; Petersen et al., 2013).

While we did not have an adequate sample size for GWAS, we were able to demonstrate that the virtual genotyped cohort design can be used successfully for both hypotheses generating research and hypothesis testing research. In this study, we were able to successfully identify both African-specific novel SNPs (Figure 6.2 and 6.4) as well as known COVID-19 and T2DM aetiological variants (Table 6.3 and 6.4). We have also demonstrated how the whole cohort may be repurposed successfully for analysis of different diseases, by the design of nested case control studies that stratify the total sample by different disease criteria. This clearly demonstrates how the cohort may be used as a disease-agnostic resource that can address many different disease outcomes. This also means the cohort design is research-agile and can be very responsive to new health challenges that arise. This agility was also demonstrated by existing population cohorts during the current COVID-19 pandemic ("Mapping the Human Genetic Architecture of COVID-19," 2021) but this work was notable in its lack of African representation (Barmania et al., 2022). As it grows over time, the cohort we are building will be able to close this gap in the future. In addition, this virtual cohort model will be even faster than traditional cohorts and health and demographic surveillance systems (HDSS) because there is no need to collect new datasets as the existing ones can be rapidly updated from the routine health data.

Through this analysis we have described how we conceptualised and implemented a genotyped virtual population cohort in a resource constrained environment. We have shown that routine health data can be effectively linked with genotyped data in a GWAS and while acknowledging the small sample size in this feasibility study, we have demonstrated that the H3Africa chip is fit for purpose and can highlight African specific variants. We are confident that this design and implementation are appropriate to scale up the cohort to a size where novel health discoveries can be made through nested case-control studies.

6.6. References

Achidi, E. A., Agbenyega, T., Allen, S., Amodu, O., Bojang, K., Conway, D., Corran, P., Deloukas, P., Djimde, A., Dolo, A., Doumbo, O., Drakeley, C., Duffy, P., Dunstan, S., Evans, J., Farrar, J., Fernando, D., Hien, T. T., Horstmann, R., … Project Management Committee. (2008). A global network for investigating the genomic epidemiology of malaria. *Nature*, *456*(7223), Article 7223. https://doi.org/10.1038/nature07632

Adebamowo, S. N., Francis, V., Tambo, E., Diallo, S. H., Landouré, G., Nembaware, V., Dareng, E., Muhamed, B., Odutola, M., Akeredolu, T., Nerima, B., Ozumba, P. J., Mbhele, S., Ghanash, A., Wachinou, A. P., & Ngomi, N. (2018). Implementation of genomics research in Africa: Challenges and recommendations. *Global Health Action*, *11*(1), 1419033. https://doi.org/10.1080/16549716.2017.1419033

Adeniyi, O. V., Yogeswaran, P., Longo-Mbenza, B., & Goon, D. T. (2016). Uncontrolled Hypertension and Its Determinants in Patients with Concomitant Type 2 Diabetes Mellitus (T2DM) in Rural South Africa. *PLOS ONE*, *11*(3), e0150033. https://doi.org/10.1371/journal.pone.0150033

Adeniyi, O. V., Yogeswaran, P., Longo-Mbenza, B., Goon, D. T., & Ajayi, A. I. (2016). Cross-sectional study of patients with type 2 diabetes in OR Tambo district, South Africa. *BMJ Open*, *6*(7), e010875. https://doi.org/10.1136/bmjopen-2015-010875

Amberbir, A., Lin, S. H., Berman, J., Muula, A., Jacoby, D., Wroe, E., Maliwichi-Nyirenda, C., Mwapasa, V., Crampin, A., Makwero, M., Singogo, E., Phiri, S., Gordon, S., Tobe, S. W., Masiye, J., Newsome, B., Hosseinipour, M., Nyirenda, M. J., & van Oosterhout, J. J. (2019). Systematic Review of Hypertension and Diabetes Burden, Risk Factors, and Interventions for Prevention and Control in Malawi: The NCD BRITE Consortium. *Global Heart*, *14*(2), 109–118. https://doi.org/10.1016/j.gheart.2019.05.001

Bertram, M. Y., Jaswal, A. V. S., Van Wyk, V. P., Levitt, N. S., & Hofman, K. J. (2013). The non-fatal disease burden caused by type 2 diabetes in South Africa, 2009. *Global Health Action*, *6*(1), 19244. https://doi.org/10.3402/gha.v6i0.19244

Bigna, J. J., & Noubiap, J. J. (2019). The rising burden of non-communicable diseases in sub-Saharan Africa. *The Lancet Global Health*, *7*(10), e1295–e1296. https://doi.org/10.1016/S2214-109X(19)30370-5

Boulle, A., Heekes, A., Tiffin, N., Smith, M., Mutemaringa, T., Zinyakatira, N., Phelanyane, F., Pienaar, C., Buddiga, K., Coetzee, E., Rooyen, R. van, Dyers, R., Fredericks, N., Loff, A., Shand, L., Moodley, M., Vega, I. de, & Vallabhjee, K. (2019). Data Centre Profile: The Provincial Health Data Centre of the Western Cape Province, South Africa. *International Journal of Population Data Science*, *4*(2), Article 2. https://doi.org/10.23889/ijpds.v4i2.1143

Brumpton, B. M., Graham, S., Surakka, I., Skogholt, A. H., Løset, M., Fritsche, L. G., Wolford, B., Zhou, W., Nielsen, J. B., Holmen, O. L., Gabrielsen, M. E., Thomas, L., Bhatta, L., Rasheed, H., Zhang, H., Kang, H. M., Hornsby, W., Moksnes, M. R., Coward, E., … Willer, C. J. (2022). The HUNT study: A population-based cohort for genetic research. *Cell Genomics*, *2*(10), 100193. https://doi.org/10.1016/j.xgen.2022.100193

Campbell, M. C., & Tishkoff, S. A. (2008). African Genetic Diversity: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping. *Annual Review of Genomics and Human Genetics*, *9*(1), 403–433. https://doi.org/10.1146/annurev.genom.9.081307.164258

Chikowore, T., Ekoru, K., Vujkovi, M., Gill, D., Pirie, F., Young, E., Sandhu, M. S., McCarthy, M., Rotimi, C., Adeyemo, A., Motala, A., & Fatumo, S. (2022). Polygenic Prediction of Type 2 Diabetes in Africa. *Diabetes Care*, *45*(3), 717–723. https://doi.org/10.2337/dc21-0365

Chimusa, E. R., Daya, M., Möller, M., Ramesar, R., Henn, B. M., Helden, P. D. van, Mulder, N. J., & Hoal, E. G. (2013). Determining Ancestry Proportions in Complex Admixture Scenarios in South Africa Using a Novel Proxy Ancestry Selection Method. *PLOS ONE*, *8*(9), e73971. https://doi.org/10.1371/journal.pone.0073971

Choudhury, A., Aron, S., Sengupta, D., Hazelhurst, S., & Ramsay, M. (2018). African genetic diversity provides novel insights into evolutionary history and local adaptations. *Human Molecular Genetics*, *27*(R2), R209–R218. https://doi.org/10.1093/hmg/ddy161

Choudhury, A., Brandenburg, J.-T., Chikowore, T., Sengupta, D., Boua, P. R., Crowther, N. J., Agongo, G., Asiki, G., Gómez-Olivé, F. X., Kisiangani, I., Maimela, E., Masemola-

Maphutha, M., Micklesfield, L. K., Nonterah, E. A., Norris, S. A., Sorgho, H., Tinto, H., Tollman, S., Graham, S. E., … Ramsay, M. (2022). Meta-analysis of sub-Saharan African studies provides insights into genetic architecture of lipid traits. *Nature Communications*, *13*(1), Article 1. https://doi.org/10.1038/s41467-022-30098-w

Choudhury, A., Ramsay, M., Hazelhurst, S., Aron, S., Bardien, S., Botha, G., Chimusa, E. R., Christoffels, A., Gamieldien, J., Sefid-Dashti, M. J., Joubert, F., Meintjes, A., Mulder, N., Ramesar, R., Rees, J., Scholtz, K., Sengupta, D., Soodyall, H., Venter, P., … Pepper, M. S. (2017). Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. *Nature Communications*, *8*(1), Article 1. https://doi.org/10.1038/s41467-017-00663-9

Choudhury, A., Sengupta, D., Ramsay, M., & Schlebusch, C. (2021). Bantu-speaker migration and admixture in southern Africa. *Human Molecular Genetics*, *30*(R1), R56–R63. https://doi.org/10.1093/hmg/ddaa274

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

Dave, J. A., Tamuhla, T., Tiffin, N., Levitt, N. S., Ross, I. L., Toet, W., Davies, M.-A., Boulle, A., Coetzee, A., & Raubenheimer, P. J. (2021). Risk factors for COVID-19 hospitalisation and death in people living with diabetes: A virtual cohort study from the Western Cape Province, South Africa. *Diabetes Research and Clinical Practice*, *177*, 108925. https://doi.org/10.1016/j.diabres.2021.108925

de Wit, E., Delport, W., Rugamika, C. E., Meintjes, A., Möller, M., van Helden, P. D., Seoighe, C., & Hoal, E. G. (2010). Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape. *Human Genetics*, *128*(2), 145–153. https://doi.org/10.1007/s00439-010-0836-1

Dissecting the phenotype in genome-wide association studies of psychiatric illness. (2009). *The British Journal of Psychiatry : The Journal of Mental Science*, *195*(2), 97–99. https://doi.org/10.1192/bjp.bp.108.063156

du Bruyn, E., Stek, C., Daroowala, R., Said-Hartley, Q., Hsiao, M., Schafer, G., Goliath, R. T., Abrahams, F., Jackson, A., Wasserman, S., Allwood, B. W., Davis, A. G., Lai, R. P.-J., Coussens, A. K., Wilkinson, K. A., de Vries, J., Tiffin, N., Cerrone, M., Ntusi, N. A. B., … Wilkinson, R. J. (2023). Effects of tuberculosis and/or HIV-1 infection on COVID-19 presentation and immune response in Africa. *Nature Communications*, *14*, 188. https://doi.org/10.1038/s41467-022-35689-1

Durbin, R. (2014). Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT). *Bioinformatics*, *30*(9), 1266–1272. https://doi.org/10.1093/bioinformatics/btu014

Forgetta, V., Li, R., Darmond-Zwaig, C., Belisle, A., Balion, C., Roshandel, D., Wolfson, C., Lettre, G., Pare, G., Paterson, A. D., Griffith, L. E., Verschoor, C., Lathrop, M., Kirkland, S., Raina, P., Richards, J. B., & Ragoussis, J. (2022). Cohort profile: Genomic data for 26 622 individuals from the Canadian Longitudinal Study on Aging (CLSA). *BMJ Open*, *12*(3), e059021. https://doi.org/10.1136/bmjopen-2021-059021

Gill, G. V., Mbanya, J.-C., Ramaiya, K. L., & Tesfaye, S. (2008). A sub-Saharan African perspective of diabetes. *Diabetologia*, *52*(1), 8. https://doi.org/10.1007/s00125-008-1167-9

Gouda, H. N., Charlson, F., Sorsdahl, K., Ahmadzada, S., Ferrari, A. J., Erskine, H., Leung, J., Santamauro, D., Lund, C., Aminde, L. N., Mayosi, B. M., Kengne, A. P., Harris, M., Achoki, T., Wiysonge, C. S., Stein, D. J., & Whiteford, H. (2019). Burden of non-communicable diseases in sub-Saharan Africa, 1990–2017: Results from the Global Burden of Disease Study 2017. *The Lancet Global Health*, *7*(10), e1375–e1387. https://doi.org/10.1016/S2214-109X(19)30374-2

H3Africa Chip. (n.d.). H3ABioNet - Pan African Bioinformatics Network. Retrieved February 6, 2023, from https://www.h3abionet.org/h3africa-chip

Hall, V., Thomsen, R. W., Henriksen, O., & Lohse, N. (2011). Diabetes in Sub Saharan Africa 1999-2011: Epidemiology and public health implications. a systematic review. *BMC Public Health*, *11*(1), 564. https://doi.org/10.1186/1471-2458-11-564

Hewitt, J., Walters, M., Padmanabhan, S., & Dawson, J. (2016). Cohort profile of the UK Biobank: Diagnosis and characteristics of cerebrovascular disease. *BMJ Open*, *6*(3), e009161. https://doi.org/10.1136/bmjopen-2015-009161

Johnston, H. R., Hu, Y.-J., Gao, J., O'Connor, T. D., Abecasis, G. R., Wojcik, G. L., Gignoux, C. R., Gourraud, P.-A., Lizee, A., Hansen, M., Genuario, R., Bullis, D., Lawley, C., Kenny, E. E., Bustamante, C., Beaty, T. H., Mathias, R. A., Barnes, K. C., & Qin, Z. S. (2017). Identifying tagging SNPs for African specific genetic variation from the African Diaspora Genome. *Scientific Reports*, *7*(1), Article 1. https://doi.org/10.1038/srep46398

Kamiza, A. B., Toure, S. M., Vujkovic, M., Machipisa, T., Soremekun, O. S., Kintu, C., Corpas, M., Pirie, F., Young, E., Gill, D., Sandhu, M. S., Kaleebu, P., Nyirenda, M., Motala, A. A., Chikowore, T., & Fatumo, S. (2022). Transferability of genetic risk scores in African populations. *Nature Medicine*, *28*(6), 1163–1166. https://doi.org/10.1038/s41591-022-01835-x

Lemma, S., Janson, A., Persson, L.-Å., Wickremasinghe, D., & Källestål, C. (2020). Improving quality and use of routine health information system data in low- and middle-income countries: A scoping review. *PLOS ONE*, *15*(10), e0239683. https://doi.org/10.1371/journal.pone.0239683

Levitt, N. S. (2008). Diabetes in Africa: Epidemiology, management and healthcare challenges. *Heart*, *94*(11), 1376–1382. https://doi.org/10.1136/hrt.2008.147306

Loh, P.-R., Danecek, P., Palamara, P. F., Fuchsberger, C., A Reshef, Y., K Finucane, H., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G. R., Durbin, R., & L Price, A. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics*, *48*(11), Article 11. https://doi.org/10.1038/ng.3679

Manyema, M., Veerman, J. L., Chola, L., Tugendhaft, A., Labadarios, D., & Hofman, K. (2015). Decreasing the Burden of Type 2 Diabetes in South Africa: The Impact of Taxing Sugar-Sweetened Beverages. *PLOS ONE*, *10*(11), e0143050. https://doi.org/10.1371/journal.pone.0143050

Marees, A. T., Kluiver, H. de, Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., & Derks, E. M. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research*, *27*(2). https://doi.org/10.1002/mpr.1608

Matimba, A., Oluka, M. N., Ebeshi, B. U., Sayi, J., Bolaji, O. O., Guantai, A. N., & Masimirembwa, C. M. (2008). Establishment of a biobank and pharmacogenetics database of African populations. *European Journal of Human Genetics*, *16*(7), Article 7. https://doi.org/10.1038/ejhg.2008.49

May, A., Hazelhurst, S., Li, Y., Norris, S. A., Govind, N., Tikly, M., Hon, C., Johnson, K. J., Hartmann, N., Staedtler, F., & Ramsay, M. (2013). Genetic diversity in black South Africans from Soweto. *BMC Genomics*, *14*(1), 644. https://doi.org/10.1186/1471-2164-14-644

Mulder, N. (2017). Development to enable precision medicine in Africa. *Personalized Medicine*, *14*(6), 467–470. https://doi.org/10.2217/pme-2017-0055

Mulder, N., Abimiku, A., Adebamowo, S. N., de Vries, J., Matimba, A., Olowoyo, P., Ramsay, M., Skelton, M., & Stein, D. J. (2018). H3Africa: Current perspectives. *Pharmacogenomics and Personalized Medicine*, *11*, 59–66. https://doi.org/10.2147/PGPM.S141546

Mulder, N. J., Adebiyi, E., Adebiyi, M., Adeyemi, S., Ahmed, A., Ahmed, R., Akanle, B., Alibi, M., Armstrong, D. L., Aron, S., Ashano, E., Baichoo, S., Benkahla, A., Brown, D. K., Chimusa, E. R., Fadlelmola, F. M., Falola, D., Fatumo, S., Ghedira, K., … H3ABioNet Consortium, as members of the H3Africa Consortium. (2017). Development of Bioinformatics Infrastructure for Genomics Research. *Global Heart*, *12*(2), 91–98. https://doi.org/10.1016/j.gheart.2017.01.005

Nagai, A., Hirata, M., Kamatani, Y., Muto, K., Matsuda, K., Kiyohara, Y., Ninomiya, T., Tamakoshi, A., Yamagata, Z., Mushiroda, T., Murakami, Y., Yuji, K., Furukawa, Y., Zembutsu, H., Tanaka, T., Ohnishi, Y., Nakamura, Y., & Kubo, M. (2017). Overview of the BioBank Japan Project: Study design and profile. *Journal of Epidemiology*, *27*(3 Suppl), S2–S8. https://doi.org/10.1016/j.je.2016.12.005

Nembaware, V., Johnston, K., Diallo, A. A., Kotze, M. J., Matimba, A., Moodley, K., Tangwa, G. B., Torrorey-Sawe, R., & Tiffin, N. (2019). A framework for tiered informed consent for health genomic research in Africa. *Nature Genetics*, *51*(11), 1566–1571. https://doi.org/10.1038/s41588-019-0520-x

Petersen, D. C., Libiger, O., Tindall, E. A., Hardie, R.-A., Hannick, L. I., Glashoff, R. H., Mukerji, M., Consortium, I. G. V., Fernandez, P., Haacke, W., Schork, N. J., & Hayes, V. M. (2013). Complex Patterns of Genomic Admixture within Southern Africa. *PLOS Genetics*, *9*(3), e1003309. https://doi.org/10.1371/journal.pgen.1003309

Policy Paper: A Framework for the Implementation of Genomic Medicine for Public Health in Africa | The AAS. (n.d.). Retrieved January 25, 2023, from https://www.aasciences.africa/publications/policy-paper-framework-implementation-genomic-medicine-public-health-africa

Popejoy, A. B., & Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature*, *538*(7624), Article 7624. https://doi.org/10.1038/538161a

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *American Journal of Human Genetics*, *81*(3), 559–575.

Ramsay, M. (2012). Africa: Continent of genome contrasts with implications for biomedical research and health. *FEBS Letters*, *586*(18), 2813–2819. https://doi.org/10.1016/j.febslet.2012.07.061

Sanger Imputation Service—Wellcome Sanger Institute. (n.d.). Retrieved February 6, 2023, from https://www.sanger.ac.uk/tool/sanger-imputation-service/

Tamuhla, T., Dave, J. A., Raubenheimer, P., & Tiffin, N. (2021). Diabetes in a TB and HIV-endemic South African population: Analysis of a virtual cohort using routine health data. *PLOS ONE*, *16*(5), e0251303. https://doi.org/10.1371/journal.pone.0251303

Tamuhla, T., Tiffin, N., & Allie, T. (2022). An e-consent framework for tiered informed consent for human genomic research in the global south, implemented as a REDCap template. *BMC Medical Ethics*, *23*(1), 119. https://doi.org/10.1186/s12910-022-00860-2

Team, R. C. (2020). *R: A Language and Environment for Statistical Computing*.

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *1*(1), Article 1. https://doi.org/10.1038/s43586-021-00056-9

Use of glycated haemoglobin (HbA1c) in the diagnosis of diabetes mellitus. (2011). *Diabetes Research and Clinical Practice*, *93*(3), 299–309. https://doi.org/10.1016/j.diabres.2011.03.012

Wabiri, N., Naidoo, I., Mungai, E., Samuel, C., & Ngwenya, T. (2019). The Arts and Tools for Using Routine Health Data to Establish HIV High Burden Areas: The Pilot Case of KwaZulu-Natal South Africa. *Frontiers in Public Health*, *7*. https://www.frontiersin.org/articles/10.3389/fpubh.2019.00335

7. Chapter 7: Discussion and conclusion

7.1 Introduction

The overarching aim of this research was to describe and pilot a pragmatic study design that would facilitate the implementation of a cost-effective scalable virtual African genotyped cohort that can be used to identify disease causing variants in African populations. This is because despite recent efforts to increase the amount of African genomic data being generated (Achidi et al., 2008; Choudhury et al., 2017; N. Mulder et al., 2018; N. J. Mulder et al., 2017), there is still a marked under-representation of African populations in genomic research (Campbell & Tishkoff, 2008; Popejoy & Fullerton, 2016) which is largely attributable to the prohibitive cost of doing human genomic research in Africa. In addressing the prohibitive cost of doing human genomic research in Africa, in this study, we have demonstrated that a virtual genotype cohort is an economically viable option for large scale genomic research in low resource settings.

In the next sections, a synopsis of key findings, strengths and limitations of the study, and future work will be presented.

7.2 Summary of key findings

Setting up a robust informed consent process

Creating a virtual genotyped cohort involves asking for sensitive genomic data and access to health records, therefore special attention should be given to interactions with the participants and the consent process, to ensure that any consent given is truly informed especially in genomic research where there is a privacy risk even with deidentified data (Tiffin, 2019). For the genotyped cohort we identified the tiered informed consent model as the most suitable for conducting human genomic research in African populations and optimised it as an electronic tiered informed consent (e-consent) framework based in REDCap (Frelich et al., 2015).

We designed the framework as a modular template that can be downloaded from GitHub (https://github.com/CIDRI-Africa/e-Consent-framework) thus providing researchers who are setting up genomic research studies with a ready-to-use tool that can be easily adapted to their research needs. Additionally, because the e-consent is based in REDCap, it allows direct data capture without the need for transcription from paper to database addressing the issue of long-term storage of paper-based consent forms. This feature not only helps to improve efficiency in collecting and storing study data, but it also allows for ease of collaboration and ethical sharing of data and biospecimens. This is because the tiered consent model we proposed offers an option for re-contact where consenting participants can be contacted for future studies. This option for re-contact also works to reduce participant research burden because no follow-up visits are required. For the virtual genotyped cohort, this functionality will make it easy to grow and follow-up the cohort over time.

Longitudinal routine health data in research

Data collection in health research is a resource intensive process and secondary use of routine health data is a cost-effective alternative because they are a rich source of population level data that can be used to describe complex disease outcomes (Gavrielov-Yusim & Friger, 2014; Grath-Lone et al., 2022; Mazzali & Duca, 2015; Yu et al., 2016). To harness this utility, we opted to use longitudinal routine health data from the PHDC for the genotyped cohort. The PHDC data contain patient demographics, laboratory results, prescribed medications and hospital encounters which are primarily collected for continuation of patient care (Boulle et al., 2019). In Chapters 3, 4 and 5 we demonstrate how we successfully re-purposed these data (Grath-Lone et al., 2022) for the virtual genotyped cohort.

The primary aim of the epidemiologic analysis in chapter 3 was to use routine health data to describe the population from which the genotyped participants would be recruited. Using an HIV comorbidities cohort data set from the PHDC, we generated summary statistics from the demographic, laboratory and pharmacy data and used them describe the epidemiology of diabetes in a virtual cohort of 16979 PLWD who have a high burden of HIV and TB. Since the data were longitudinal, we described the relationship between previous TB, newly diagnosed TB disease and HIV infection on diabetes using HbA1c results as an outcome measure. Our findings showed that 13% of the study population had a history of active TB disease and 18% were HIV positive. The HIV positive group had diabetes ascertained at a significantly younger age corroborating similar findings reported by Osei-Yeboah and colleagues (Osei-Yeboah et al., 2021) and had significantly higher HbA1c values. There was no evidence of TB disease influencing the trajectory of glycaemic control in the long term, but diabetes patients who developed active TB had higher mortality than those without TB. The epidemiologic findings in this exploratory study demonstrated that routine health data are a valuable resource for understanding disease epidemiology in resource limited settings.

For the epidemiologic analysis in Chapter 4, we were able to demonstrate that routine health data are research-agile and can be used to describe new health challenges that arise. Using the data analysis pipelines from Chapter 3 we were able to describe the epidemiology of COVID-19 in PLWD in the Western Cape Province, using data from the first wave of the pandemic. We showed that PLWD were significantly more likely to be admitted from COVID-19 compared to those without the diabetes and that HIV infection, chronic kidney disease, current TB, male sex and increasing age were all significant risk factors for severe COVID-19 in diabetes patient. While CKD, male sex, HIV infection, previous TB and increasing age were also significant risk factors for death. Pre-infection use of insulin was associated with a significant increased risk for hospitalisation and mortality and metformin was associated with a reduced risk for hospitalisation and mortality. In addition, we optimise standardised methods for defining disease phenotypes from routine health data and for the genotyped cohort, these methods will be crucial in defining accurate reproducible phenotypes for genomic association studies.

In Chapter 5 we demonstrated how routine health data can be used to identify complex phenotypes such as response to medication, long term outcomes and changes in health status. We used the pharmacy dispensing data the PHDC from to describe longitudinal persistence and adherence patterns to oral diabetes medication and determine predictors of longitudinal adherence in PLWD. Using a k-means clustering algorithm we were able to cluster individuals into four longitudinal medication adherence trajectories which described their medication use patters. These trajectories were We also investigated factors that determine long-term adherence in our study population which could be used for targeted interventions. Our results showed that PLWD on ART were more likely to have long term adherence. This might be because HIV infection in particular adherence to ART is well managed in South Africa therefore, PLWHIV are accessing health care more (Osei-Yeboah et al., 2021) and are therefore likely to have better linkage to care for other comorbidities including diabetes. Given these observations, it might be worth modelling the level of care given to PLWHIV in South Africa to PLWD.

Data integration

Having developed a robust tiered informed process and optimised methods to generate accurate and reproducible disease phenotypes form routine health data, in Chapter 6 we conducted a feasibility study piloting the implementation of a virtual genotyped cohort. Our results showed that the cohort study design we have proposed is achievable in resource limited settings. This is because a DNA sample is only collected once from a participant to generate the genotype data meaning that there is no need to set up and maintain infrastructure such as biobanks (Brumpton et al., 2022; Matimba et al., 2008; Nagai et al., 2017) for the storage of large sample collections. The tiered consent model we proposed in Chapter 2 (Tamuhla et al., 2022) also offers an option for re-contact where consenting participants can be contacted for future studies which might need additional sample collection.

Additionally, by having indefinite virtual follow-up through routine health data access, we have demonstrated that nested case-control studies can be done where consenting participants in the genotyped cohort can be cases or controls depending on the outcome being measured therefore making it possible to study multiple health conditions without the added cost of recruiting participants each time. Additionally, the continued update of routine health data over the life course means that in nested case control studies you can retrospectively correct for ascertainment bias as people who were designated as controls in initial studies may have later become cases as they develop specific phenotypes of interest.

7.3. Strengths and limitations

The WCGDH has one of the most mature health informatics environments in the region. This is also facilitated by having established a consistent unique health identifier from inception of digital health platforms in the province which has made it possible to collate data in a health exchange. Whilst other countries have nascent health informatics infrastructure and digital health records such as OpenMRS (Muhoza et al., 2019) in Rwanda and SmartCare (Kaumba, 2023) Zambia. The mature system in the WC has made it possible for the PHDC to be implemented and it is the first health information exchange of its kind in Africa.

Through this research we have shown that a virtual genotyped cohort is a pragmatic model that works in resource limited settings. This is because it supports a virtuous cycle whereby investing in development of routine health care delivery electronic data platforms such as the PHDC can both improve patient outcomes through better continuity of care and through better evidence-based public health; whilst simultaneously supporting research to generate that evidence in an environment that does not over burden both the health system and study participants with parallel data and sample collection environments.

In summary the strengths of this cohort design are:

- There is no need for storage of large/complex sample collections in biobanks as there is an option to recontact participants for future studies where more complex samples and data might be needed.
- The virtual genotyped cohort design can be used successfully for both hypothesis generating and hypothesis testing research.
- The virtual cohort can be used as a disease-agnostic resource that makes the study of all health conditions possible.

- The cohort design is research-agile and can be very responsive to new health challenges that arise.
- The cohort can be grown over time and offers a collaborative participant recruitment strategy where relevant consents are in place.
- The design offers indefinite follow up through routine health data access thus creating an open ended longitudinal virtual cohort.
- Because we ask for permission to re-contact participants, we are also able to use the cohort to design future detailed studies where we can pre-select consenting participants based on their health and genomic data and invite them to participate in future more targeted studies. As the size of cohort increases this will become feasible even if substantial number of participants cannot be contacted with existing details of decline to participate.

We have chosen to incorporate the questions together with the text in the informed consent process. We recognize that this may be a limitation in certain scenarios where IRBs may require the traditional format where questions are posed at the end of providing the information. This could be addressed in future versions by separating the elements containing the text and associated consent questions so that these can be arranged independently as required.

The biggest limitation of this cohort design is that it is based on the availability of electronic routine health data platform such as the PHDC in the Western Cape Province. A major limitation of routine health data is that they are limited to administrative data and disease states can only be inferred from the available data. They also do not contain socio-economic data which are important confounding factors. Additionally, the data are mainly used for exploratory analysis and targeted studies are required to confirm the findings.

7.4. Future work

We have been able to create a virtual genotyped cohort of 343 individuals from the Western Cape Province, South Africa and we believe the results from this work provide a strong motivation for expanding such efforts and we intend to continue building this cohort in the future. As the cohort grows, we will do more complex statistical analysis like latent factor analysis and mendelian randomisation, longitudinal data analysis including regression discontinuity. In addition, while we are currently only looking at genotype data as they are cheaper to generate, we hope in the future as whole genome sequencing (WGS) costs drop and local capacity to perform WGS increases we will be able to do WGS on all or at least a subset of our data. Finally, as part of benefit sharing for participants in genomic research where there is currently little to no benefit to participating in the research, in the future we plan on establishing a mechanism where clinically actionable results can be returned into the health system.

7.5. References

Achidi, E. A., Agbenyega, T., Allen, S., Amodu, O., Bojang, K., Conway, D., Corran, P., Deloukas, P., Djimde, A., Dolo, A., Doumbo, O., Drakeley, C., Duffy, P., Dunstan, S., Evans, J., Farrar, J., Fernando, D., Hien, T. T., Horstmann, R., … Project Management Committee. (2008). A global network for investigating the genomic epidemiology of malaria. *Nature*, *456*(7223), Article 7223. https://doi.org/10.1038/nature07632

Boulle, A., Heekes, A., Tiffin, N., Smith, M., Mutemaringa, T., Zinyakatira, N., Phelanyane, F., Pienaar, C., Buddiga, K., Coetzee, E., Rooyen, R. van, Dyers, R., Fredericks, N., Loff, A., Shand, L., Moodley, M., Vega, I. de, & Vallabhjee, K. (2019). Data Centre Profile: The Provincial Health Data Centre of the Western Cape Province, South Africa. *International Journal of Population Data Science*, *4*(2), Article 2. https://doi.org/10.23889/ijpds.v4i2.1143

Brumpton, B. M., Graham, S., Surakka, I., Skogholt, A. H., Løset, M., Fritsche, L. G., Wolford, B., Zhou, W., Nielsen, J. B., Holmen, O. L., Gabrielsen, M. E., Thomas, L., Bhatta, L., Rasheed, H., Zhang, H., Kang, H. M., Hornsby, W., Moksnes, M. R., Coward, E., … Willer, C. J. (2022). The HUNT study: A population-based cohort for genetic research. *Cell Genomics*, *2*(10), 100193. https://doi.org/10.1016/j.xgen.2022.100193

Campbell, M. C., & Tishkoff, S. A. (2008). African Genetic Diversity: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping. *Annual Review of Genomics and Human Genetics*, *9*(1), 403–433. https://doi.org/10.1146/annurev.genom.9.081307.164258

Choudhury, A., Ramsay, M., Hazelhurst, S., Aron, S., Bardien, S., Botha, G., Chimusa, E. R., Christoffels, A., Gamieldien, J., Sefid-Dashti, M. J., Joubert, F., Meintjes, A., Mulder, N., Ramesar, R., Rees, J., Scholtz, K., Sengupta, D., Soodyall, H., Venter, P., … Pepper, M. S. (2017). Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. *Nature Communications*, *8*(1), Article 1. https://doi.org/10.1038/s41467-017-00663-9

Frelich, M. J., Bosler, M. E., & Gould, J. C. (2015). Research Electronic Data Capture (REDCap) electronic Informed Consent Form (eICF) is compliant and feasible in a clinical research setting. *International Journal of Clinical Trials*, *2*(3), 51–55. https://doi.org/10.18203/2349-3259.ijct20150591

Gavrielov-Yusim, N., & Friger, M. (2014). Use of administrative medical databases in population-based research. *J Epidemiol Community Health*, *68*(3), 283–287. https://doi.org/10.1136/jech-2013-202744

Grath-Lone, L. M., Jay, M. A., Blackburn, R., Gordon, E., Zylbersztejn, A., Wiljaars, L., & Gilbert, R. (2022). What makes administrative data research-ready? : A systematic review and thematic analysis of published literature. *International Journal of Population Data Science*, *7*(1), Article 1. https://doi.org/10.23889/ijpds.v7i1.1718

Matimba, A., Oluka, M. N., Ebeshi, B. U., Sayi, J., Bolaji, O. O., Guantai, A. N., & Masimirembwa, C. M. (2008). Establishment of a biobank and pharmacogenetics database of African populations. *European Journal of Human Genetics*, *16*(7), Article 7. https://doi.org/10.1038/ejhg.2008.49

Mazzali, C., & Duca, P. (2015). Use of administrative data in healthcare research. *Internal and Emergency Medicine*, *10*(4), 517–524. https://doi.org/10.1007/s11739-015-1213-9

Mulder, N., Abimiku, A., Adebamowo, S. N., de Vries, J., Matimba, A., Olowoyo, P., Ramsay, M., Skelton, M., & Stein, D. J. (2018). H3Africa: Current perspectives. *Pharmacogenomics and Personalized Medicine*, *11*, 59–66. https://doi.org/10.2147/PGPM.S141546

Mulder, N. J., Adebiyi, E., Adebiyi, M., Adeyemi, S., Ahmed, A., Ahmed, R., Akanle, B., Alibi, M., Armstrong, D. L., Aron, S., Ashano, E., Baichoo, S., Benkahla, A., Brown, D. K., Chimusa, E. R., Fadlelmola, F. M., Falola, D., Fatumo, S., Ghedira, K., … H3ABioNet Consortium, as members of the H3Africa Consortium. (2017). Development of Bioinformatics Infrastructure for Genomics Research. *Global Heart*, *12*(2), 91–98. https://doi.org/10.1016/j.gheart.2017.01.005

Nagai, A., Hirata, M., Kamatani, Y., Muto, K., Matsuda, K., Kiyohara, Y., Ninomiya, T., Tamakoshi, A., Yamagata, Z., Mushiroda, T., Murakami, Y., Yuji, K., Furukawa, Y., Zembutsu, H., Tanaka, T., Ohnishi, Y., Nakamura, Y., & Kubo, M. (2017). Overview of the BioBank Japan Project: Study design and profile. *Journal of Epidemiology*, *27*(3 Suppl), S2–S8. https://doi.org/10.1016/j.je.2016.12.005

Osei-Yeboah, R., Tamuhla, T., Ngwenya, O., & Tiffin, N. (2021). Accessing HIV care may lead to earlier ascertainment of comorbidities in health care clients in Khayelitsha, Cape Town. *PLOS Global Public Health*, *1*(12), e0000031. https://doi.org/10.1371/journal.pgph.0000031

Popejoy, A. B., & Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature*, *538*(7624), Article 7624. https://doi.org/10.1038/538161a

Tamuhla, T., Tiffin, N., & Allie, T. (2022). An e-consent framework for tiered informed consent for human genomic research in the global south, implemented as a REDCap template. *BMC Medical Ethics*, *23*(1), 119. https://doi.org/10.1186/s12910-022-00860-2

Tiffin, N. (2019). Potential risks and solutions for sharing genome summary data from African populations. *BMC Medical Genomics*, *12*(1), 152. https://doi.org/10.1186/s12920-019-0604-6

Yu, A. Y. X., Holodinsky, J. K., Zerna, C., Svenson, L. W., Jetté, N., Quan, H., & Hill, M. D. (2016). Use and Utility of Administrative Health Data for Stroke Research and Surveillance. *Stroke*, *47*(7), 1946–1952. https://doi.org/10.1161/STROKEAHA.116.012390

Appendix : Supplementary files

Supplementary file 2.1

Participant information and informed consent checklist for a new research study.

1. **Steps to create the tiered consent workflow using the REDCap template:**

    1.1. Set up research study database in REDCap
    1.2. Download tiered e-consent template codebook (ConsentFramework_Data_Dictionary) and supporting documents from GitHub repository
    1.3. Import tiered e-consent template codebook in REDCap
    1.4. Use the guidance documents provided to set up and enable e-consent module in REDCap
    1.5. Submit e-consent documents to relevant institutional review board for ethics approval
    1.6. Train research staff on administering tiered e-consent
    1.7. Implement use of e-consent in new research study participant recruitment

2. **Participant information and informed consent modules to include:**

Type of consent	✓/ ×
Primary consent for collecting biospecimens and health data for specific disease in current study.	

Consent for access to medical records	
Consent for return of individual results	
Consent for return of individual results that are actionable and/or treatable	
Consent for return of individual results that are NOT actionable and/or treatable	
Consent for inclusion of individual data in genetic summary data	
Consent for use of genetic and health data for future studies on specific disease	
Consent for use of genetic and health data for future studies on other health conditions or related health processes	
Consent to re-contact for future studies	
Consent for use of genetic and health data in international studies	
Consent for use of genetic data in population origins and ancestry studies	

Supplementary file 2.2

Supplementary Table 2: List of the documents in the tiered e-consent framework Github repository.

Name in GitHub Repository (including link to document)	**Document content**	**Use of content**
ConsentFramework.xml https://github.com/CIDRI-Africa/e-Consent-framework/blob/main/ConsentFramework_2021-09-29_1108.REDCap.xml	XML file which contains the entire tiered e-consent project metadata	This can be imported into REDCap to set up a clone of the project.
ConsentFramework_Data_Dictionary https://github.com/CIDRI-Africa/e-Consent-framework/blob/main/ConsentFramework_DataDictionary_2021-09-29.csv	CSV file which contains all the tiered e-consent variables	This is the codebook that can be used to set up a new instance of the e-consent framework
ConsentFramework_All_Documents https://github.com/CIDRI-Africa/e-Consent-framework/blob/main/ConsentFramework_Allforms_20210929.pdf	PDF of example copies of the different documents generated used the tiered e-consent framework	Example output

Instrument index.xls https://github.com/CIDRI-Africa/e-Consent-framework/blob/main/Instrument%20index%2020210929.xlsx	List of all data capture instruments available in the e-consent	Shows which instruments are essential and which are optional when implementing your own instance of the framework
Set up guide.doc https://github.com/CIDRI-Africa/e-Consent-framework/blob/main/Set-up%20guide%2020200204.docx	A word document	Step by step guide on how to set up REDCap and use the tiered e-consent feature
Diabetes study example.pdf https://github.com/CIDRI-Africa/e-Consent-framework/blob/main/Diabetes_study_example.pdf	A PDF document	An example of a tiered e-consent document showing the final output from the different data capture tools.

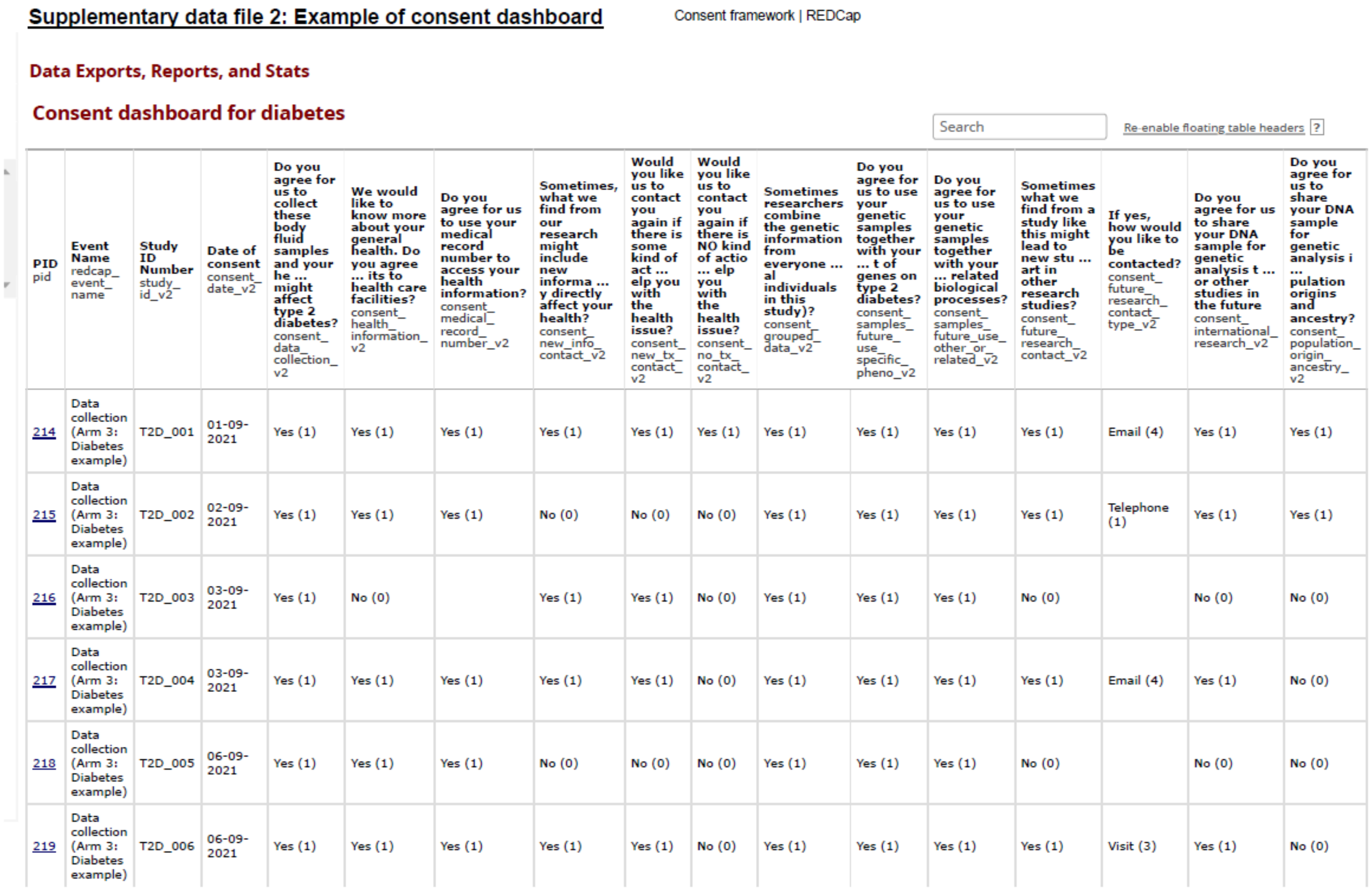
Supplementary file 2.3

Supplementary Table 1: Additional REDCap survey customisations that were used in the tiered e-consent documents

Customisation	**Utility of customisation**
Set a Custom Record Label	Allows another variable to the appended to the system generated record name to aid in ease of identification of individual participants records
Designate a Secondary Unique Field	A unique constraint value which cannot be duplicated and will be checked in real time to ensure that is not shared by another record e.g. participant study ID
Require a reason when making changes to existing records	Require users to enter a reason (200 character max) in a text box when making any data changes to an already existing record on a data collection instrument. The prompt is triggered when clicking the Save button on the page. Any 'reasons' entered can then be viewed anytime afterward on the Logging page. his feature is only triggered when adding, editing, or deleting data for an instrument that contains previously-collected data for one or more fields on the instrument.
Display the Today/Now button for all date and time fields on forms/surveys?	If enabled, a 'Today' button will be displayed to the right of all date fields, and a 'Now' button will be displayed to the right of all time, datetime,

	and datetime_seconds fields. Clicking the button will automatically set the field's value with the current date or time.
Enable the File Version History for 'File Upload' fields?	If a new version of a file needs to be uploaded for the field, instead of deleting the current file before adding the new one, you may simply upload a new file (via the 'Upload new version' link), in which all older versions will be kept and will be accessible for viewing/download in the Data History popup for the field. This features provides the convenience of accessing older versions of the file instead of having to delete them. (Note: Older versions of a file will not be accessible anywhere else in the project except the Data History popup.
Enable the Data History popup for all data collection instruments?	If enabled, an icon will appear next to every field on a data collection instrument. When the icon is clicked, the history of all data entered into that field for that record will be listed chronologically and will display all previous values, who changed the value at each instance, and the time it was changed.
Enable the Field Comment Log or Data Resolution Workflow (Data Queries)?	The Field Comment Log (enabled by default) allows users to leave comments for any given field on a data entry form by clicking the balloon icon next to the field. All comments can also be viewed, searched, and downloaded on the Field Comment Log page.

Supplementary file 2.4



Supplementary data file 2: Example of consent dashboard

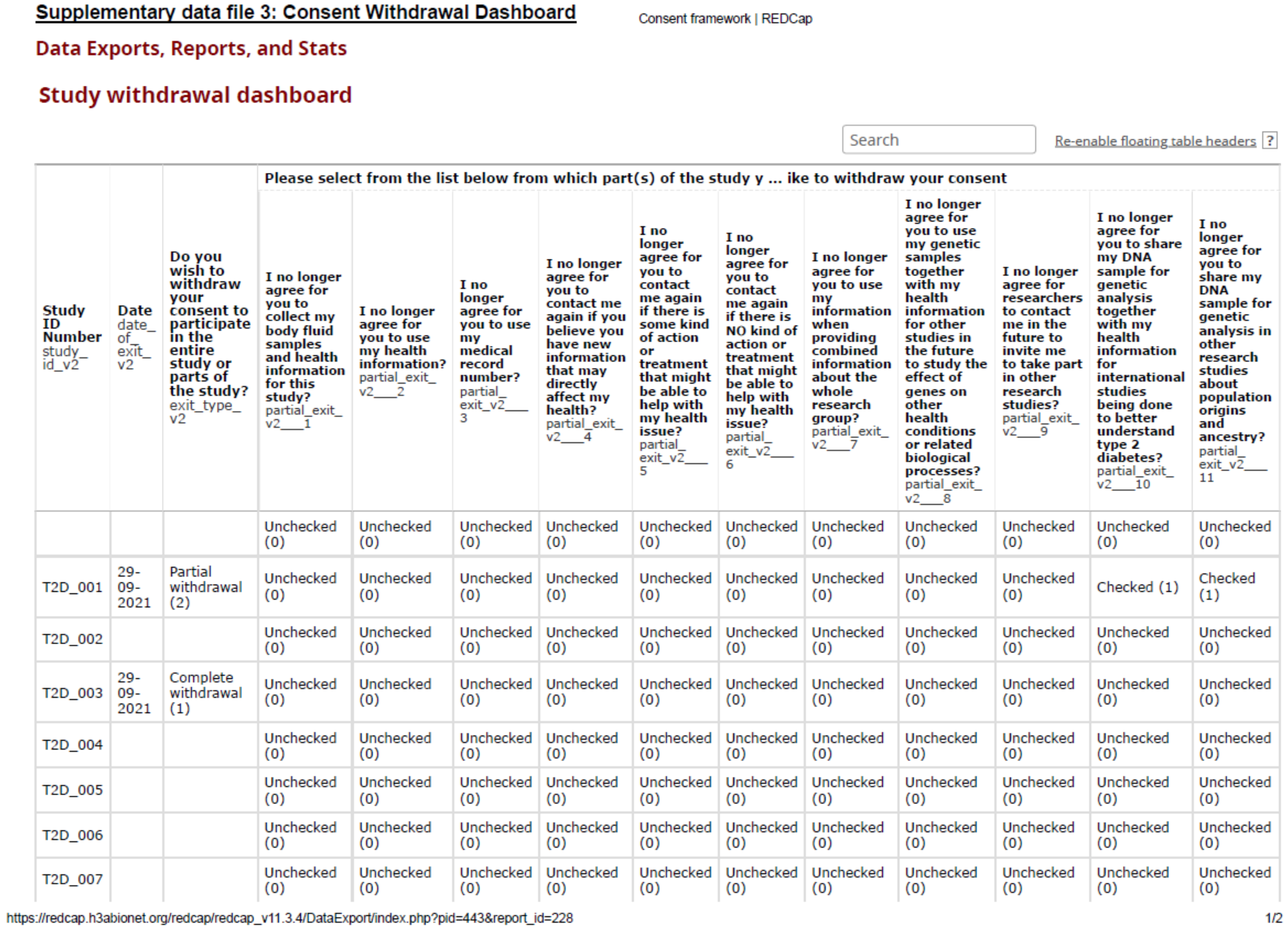
Consent framework | REDCap

Data Exports, Reports, and Stats

Consent dashboard for diabetes

PID pid	**Event Name** redcap_ event_ name	**Study ID Number** study_ id_v2	**Date of consent** consent_ date_v2	**Do you agree for us to collect these body fluid samples and your he ... might affect type 2 diabetes?** consent_ data_ collection_ v2	**We would like to know more about your general health. Do you agree ... its to health care facilities?** consent_ health_ information_ v2	**Do you agree for us to use your medical record number to access your health information?** consent_ medical_ record_ number_v2	**Sometimes, what we find from our research might include new informa ... y directly affect your health?** consent_ new_info_ contact_v2	**Would you like us to contact you again if there is some kind of act ... elp you with the health issue?** consent_ new_tx_ contact_ v2	**Would you like us to contact you again if there is NO kind of actio ... elp you with the health issue?** consent_ no_tx_ contact_ v2	**Sometimes researchers combine the genetic information from everyone ... al individuals in this study)?** consent_ grouped_ data_v2	**Do you agree for us to use your genetic samples together with your ... t of genes on type 2 diabetes?** consent_ samples_ future_ use_ specific_ pheno_v2	**Do you agree for us to use your genetic samples together with your ... related biological processes?** consent_ samples_ future_use_ other_or_ related_v2	**Sometimes what we find from a study like this might lead to new stu ... art in other research studies?** consent_ future_ research_ contact_v2	**If yes, how would you like to be contacted?** consent_ future_ research_ contact_ type_v2	**Do you agree for us to share your DNA sample for genetic analysis t ... or other studies in the future** consent_ international_ research_v2	**Do you agree for us to share your DNA sample for genetic analysis i ... pulation origins and ancestry?** consent_ population_ origin_ ancestry_ v2
214	Data collection (Arm 3: Diabetes example)	T2D_001	01-09-2021	Yes (1)	Yes (1)	Yes (1)	Yes (1)	Yes (1)	Yes (1)	Yes (1)	Yes (1)	Yes (1)	Yes (1)	Email (4)	Yes (1)	Yes (1)
215	Data collection (Arm 3: Diabetes example)	T2D_002	02-09-2021	Yes (1)	Yes (1)	Yes (1)	No (0)	No (0)	No (0)	Yes (1)	Yes (1)	Yes (1)	Yes (1)	Telephone (1)	Yes (1)	Yes (1)
216	Data collection (Arm 3: Diabetes example)	T2D_003	03-09-2021	Yes (1)	No (0)		Yes (1)	Yes (1)	No (0)	Yes (1)	Yes (1)	Yes (1)	No (0)		No (0)	No (0)
217	Data collection (Arm 3: Diabetes example)	T2D_004	03-09-2021	Yes (1)	Yes (1)	Yes (1)	Yes (1)	Yes (1)	No (0)	Yes (1)	Yes (1)	Yes (1)	Yes (1)	Email (4)	Yes (1)	No (0)
218	Data collection (Arm 3: Diabetes example)	T2D_005	06-09-2021	Yes (1)	Yes (1)	Yes (1)	No (0)	No (0)	No (0)	Yes (1)	Yes (1)	Yes (1)	No (0)		No (0)	No (0)
219	Data collection (Arm 3: Diabetes example)	T2D_006	06-09-2021	Yes (1)	Yes (1)	Yes (1)	Yes (1)	Yes (1)	No (0)	Yes (1)	Yes (1)	Yes (1)	Yes (1)	Visit (3)	Yes (1)	No (0)

Supplementary file 2.5



Supplementary data file 3: Consent Withdrawal Dashboard

Consent framework | REDCap

Data Exports, Reports, and Stats

Study withdrawal dashboard

Search

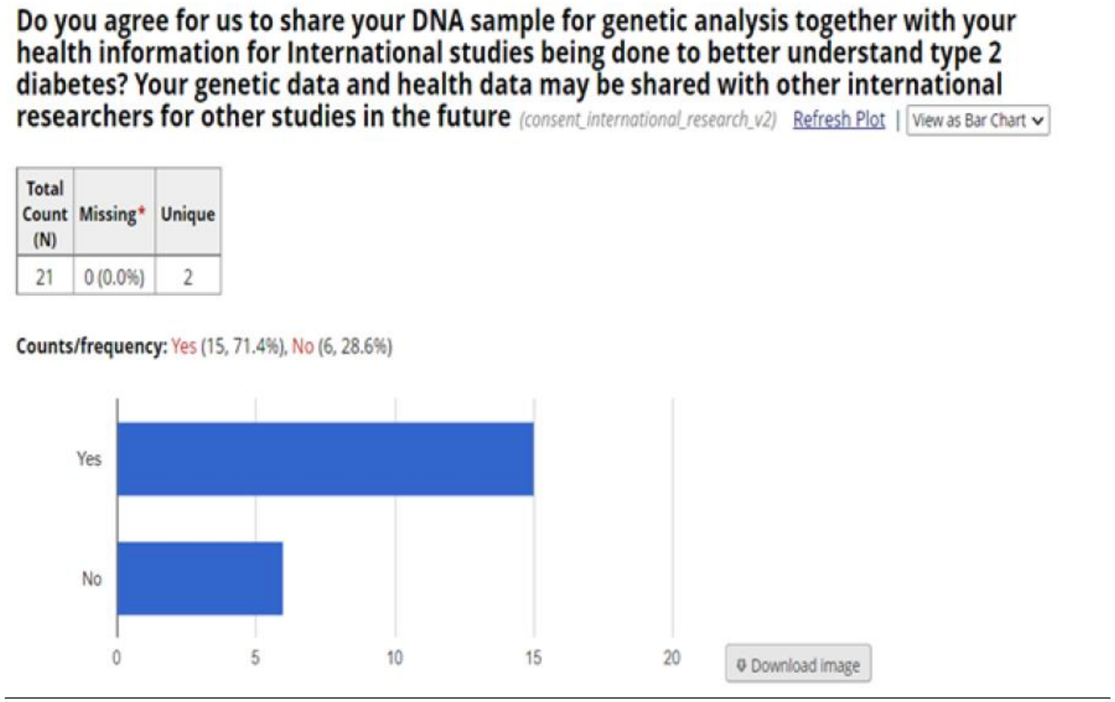
Re-enable floating table headers ?

			Please select from the list below from which part(s) of the study y ... ike to withdraw your consent										
Study ID Number study_id_v2	**Date** date_of_exit_v2	**Do you wish to withdraw your consent to participate in the entire study or parts of the study?** exit_type_v2	**I no longer agree for you to collect my body fluid samples and health information for this study?** partial_exit_v2___1	**I no longer agree for you to use my health information?** partial_exit_v2___2	**I no longer agree for you to use my medical record number?** partial_exit_v2___3	**I no longer agree for you to contact me again if you believe you have new information that may directly affect my health?** partial_exit_v2___4	**I no longer agree for you to contact me again if there is some kind of action or treatment that might be able to help with my health issue?** partial_exit_v2___5	**I no longer agree for you to contact me again if there is NO kind of action or treatment that might be able to help with my health issue?** partial_exit_v2___6	**I no longer agree for you to use my information when providing combined information about the whole research group?** partial_exit_v2___7	**I no longer agree for you to use my genetic samples together with my health information for other studies in the future to study the effect of genes on other health conditions or related biological processes?** partial_exit_v2___8	**I no longer agree for researchers to contact me in the future to invite me to take part in other research studies?** partial_exit_v2___9	**I no longer agree for you to share my DNA sample for genetic analysis together with my health information for international studies being done to better understand type 2 diabetes?** partial_exit_v2___10	**I no longer agree for you to share my DNA sample for genetic analysis in other research studies about population origins and ancestry?** partial_exit_v2___11
			Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)
T2D_001	29-09-2021	Partial withdrawal (2)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Checked (1)	Checked (1)
T2D_002			Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)
T2D_003	29-09-2021	Complete withdrawal (1)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)
T2D_004			Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)
T2D_005			Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)
T2D_006			Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)
T2D_007			Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)	Unchecked (0)

https://redcap.h3abionet.org/redcap/redcap_v11.3.4/DataExport/index.php?pid=443&report_id=228 1/2

Supplementary file 2.6

<u>**Supplementary Data File 4: Example of study population data summarised for each type of consent**</u>



Do you agree for us to share your DNA sample for genetic analysis together with your health information for International studies being done to better understand type 2 diabetes? Your genetic data and health data may be shared with other international researchers for other studies in the future *(consent_international_research_v2)* Refresh Plot | View as Bar Chart

Total Count (N)	Missing*	Unique
21	0 (0.0%)	2

Counts/frequency: Yes (15, 71.4%), No (6, 28.6%)

Supplementary Table 3.1: Characteristics of the study population who are HIV positive and stratified by the ascertainment of the HIV in relation to diabetes ascertainment

	All *N=2423[a]*	**HIV before diabetes ascertainment** *N=1525 (63%)*	**HIV after diabetes ascertainment** *N=898 (37%)*	p-value
Sex (Female)	1676 (69.2%)	1033 (67.8%)	643 (71.6%)	0.058
Age at diabetes ascertainment (Years)	46.0 [39.0;52.0]	45.0 [39.0;52.0]	47.0 [39.0;53.0]	0.029
Baseline HbA1c (%)	8.4 [6.9;10.9]	8.0 [6.8;10.4]	9.4 [7.3;11.8]	<0.001
Last HbA1c (%)	12.2 [8.0;15.0]	12.5 [7.9;15.2]	11.6 [8.0;14.5]	0.005
Patient outcome (Deceased)	142 (5.9%)	69 (4.5%)	73 (8.1%)	<0.001
Diabetes duration (Years)	4.0 [1.2;6.5]	2.9 [0.6;5.6]	5.8 [3.1;7.8]	<0.001
Ever started diabetes treatment	2056 (84.9%)	1237 (81.1%)	819 (91.2%)	<0.001
Ever had Tuberculosis	786 (32.4%)	548 (35.9%)	238 (26.5%)	<0.001
TB-Diabetes comorbidity	372 (47.8%)	164 (30.1%)	208 (88.5%)	<0.001

a. 85 (3%) individuals did not have enough data to classify when they were ascertained HIV relative to diabetes ascertainment

Supplementary Table 3.2: Characteristics of the whole study population, and stratified by history of active Tuberculosis disease

	ALL *N=13771*	**Never had Tuberculosis** *N=11932*	**Ever had Tuberculosis** *N=1839*	**p-value**
Sex (Female)	9246 (67.2%)	8201 (68.8%)	1045 (56.9%)	<0.001
Age at diabetes ascertainment (Years)	52.0 [44.0;59.0]	52.0 [44.0;60.0]	49.0 [41.0;56.0]	<0.001
Baseline HbA1c (%)	8.5 [7.0;11.1]	8.5 [7.0;11.0]	9.2 [7.1;11.8]	<0.001
Last HbA1c (%)	9.5 [7.2;12.7]	9.3 [7.1;12.5]	10.9 [7.6;14.1]	<0.001
Patient outcome (Deceased)	631 (4.6%)	448 (3.8%)	183 (10.0%)	<0.001
TB-Diabetes comorbidity	1008 (55.9%)	0 (.%)	1008 (55.9%)	.
Ever started TB treatment	1831 (13.3%)	0 (0.0%)	1831 (99.6%)	0.000
Ever started diabetes treatment	11745 (85.3%)	10141 (85.0%)	1604 (87.2%)	0.013
Linkage to diabetes Treatment	10707 (91.2%)	9335 (92.1%)	1372 (85.5%)	<0.001
Linkage to HbA1c testing	9264 (67.3%)	8029 (67.3%)	1235 (67.2%)	0.896

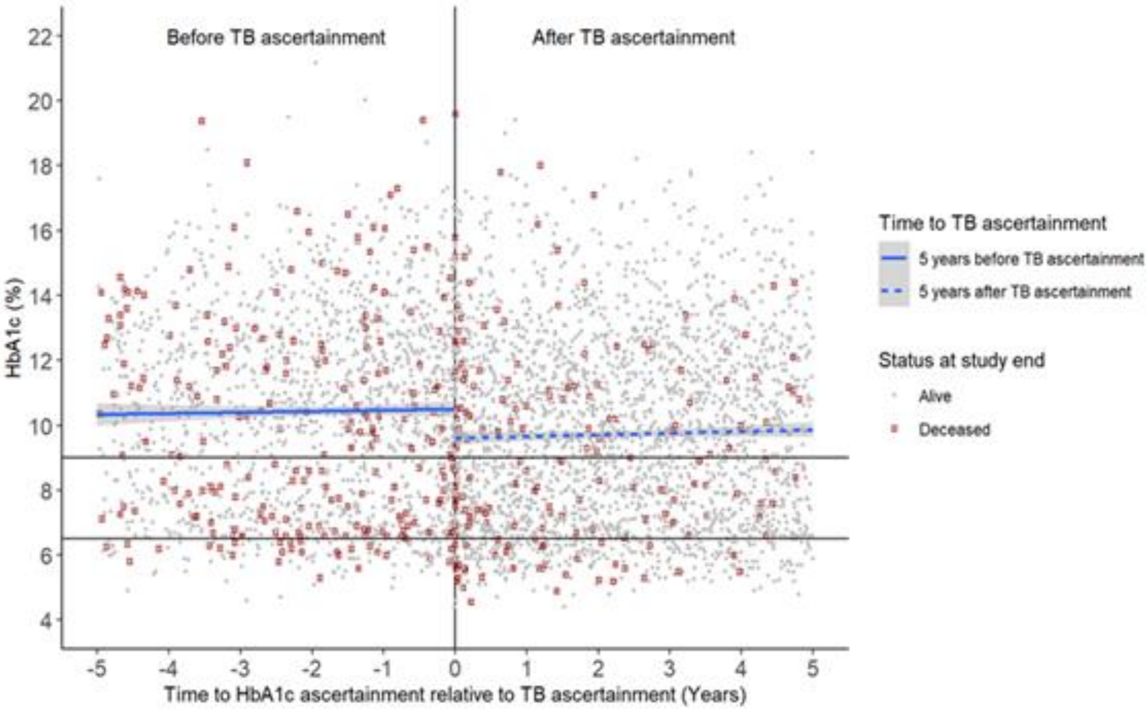
Supplementary Table 3.3: Number of study participants recorded as being on the different combinations of Diabetes, TB, and HIV medications.

Diabetes treatment	Tuberculosis treatment	HIV treatment	Number of study participants
1	0	0	10374
1	0	1	1323
1	1	0	1088
1	1	1	743

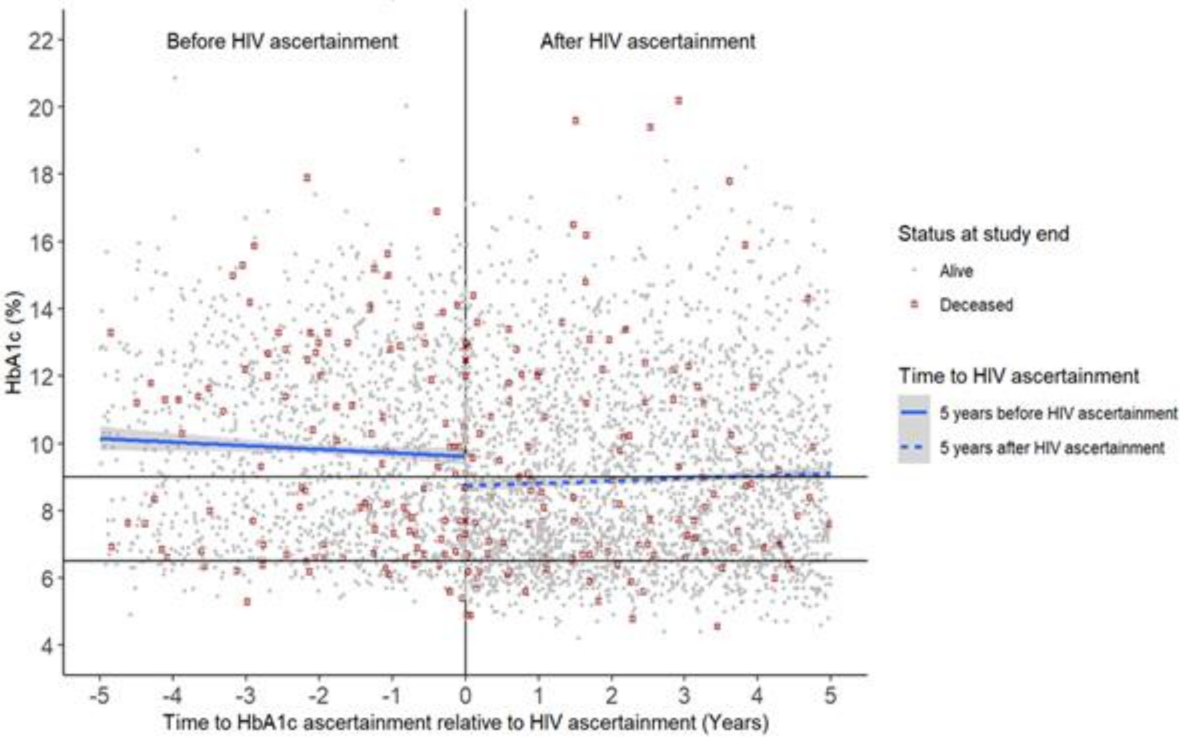
Supplementary Table 3.4: Characteristics of the whole study population, and stratified by the sex of the participants

	[ALL] *N=13754* [a]	**Female** *N=9246* (67%)	**Male** *N=4508* (33%)	**p-value**
Age at diabetes ascertainment (Years)	52.0 [44.0;59.0]	52.0 [44.0;60.0]	51.0 [43.0;59.0]	<0.001
Age categories				<0.001
18-39	2108 (15.3%)	1348 (14.6%)	760 (16.9%)	
40-49	3711 (27.0%)	2476 (26.8%)	1235 (27.4%)	
50-59	4570 (33.2%)	3090 (33.4%)	1480 (32.8%)	
60-69	2356 (17.1%)	1597 (17.3%)	759 (16.8%)	
70-79	804 (5.8%)	569 (6.2%)	235 (5.2%)	
>=80	205 (1.5%)	166 (1.8%)	39 (0.9%)	
HIV Positive	2506 (18.2%)	1726 (18.7%)	780 (17.3%)	0.054
Ever had Tuberculosis	1838 (13.4%)	1045 (11.3%)	793 (17.6%)	<0.001
Tuberculosis-Diabetes comorbidity	1007 (55.9%)	581 (56.8%)	426 (54.8%)	0.415
Baseline HbA1c (%)	8.5 [7.0;11.1]	8.6 [7.0;11.1]	8.5 [6.9;11.2]	0.441
Baseline HbA1c < 7%	2816 (24.5%)	1867 (24.0%)	949 (25.5%)	0.089
Last HbA1c (%)	9.4 [7.2;12.7]	9.5 [7.2;12.7]	9.3 [7.1;12.7]	0.038
Last HbA1c < 7%	2927 (21.3%)	1886 (20.4%)	1041 (23.1%)	<0.001
Patient outcome (Deceased)	631 (4.6%)	394 (4.3%)	237 (5.3%)	0.010
Ever started TB treatment	1830 (13.3%)	1043 (11.3%)	787 (17.5%)	<0.001
Ever started diabetes treatment	11729 (85.3%)	7802 (84.4%)	3927 (87.1%)	<0.001
Linkage to diabetes treatment	10694 (91.2%)	7117 (91.2%)	3577 (91.1%)	0.838
Linkage to HbA1c testing	9250 (67.3%)	6157 (66.6%)	3093 (68.6%)	0.021

Supplementary figure 3.1



Supplementary Figure 3.2



Supplementary Table 4.1. Characteristics of the Western Cape public health sector patients with COVID-19 who have diabetes. The results have been grouped by the type of health facility they accessed to fill drug prescriptions

	ALL *N=9305*	**Computer based Pharmacy system** *N=5721*	**No computer based pharmacy system** *N=3584*
Sex:			
Female	5645 (60.7%)	3598 (62.9%)	2047 (57.1%)
Male	3657 (39.3%)	2121 (37.1%)	1536 (42.9%)
age	55.0 [46.0;63.0]	57.0 [48.0;65.0]	52.0 [42.0;61.0]
Age:			
0-18	19 (0.2%)	9 (0.2%)	10 (0.3%)
18-39	1244 (13.4%)	539 (9.4%)	705 (19.7%)
40-49	1852 (19.9%)	1033 (18.1%)	819 (22.9%)
50-59	2886 (31.0%)	1819 (31.8%)	1067 (29.8%)
60-69	2063 (22.2%)	1437 (25.1%)	626 (17.5%)
70-79	950 (10.2%)	677 (11.8%)	273 (7.6%)
>=80	291 (3.1%)	207 (3.6%)	84 (2.3%)
Outcome:			
Active	1 (<0.1%)	0 (0.0%)	1 (<0.1%)
Died	1434 (15.4%)	928 (16.3%)	506 (14.1%)
Recovered	7856 (84.6%)	4780 (83.7%)	3076 (85.8%)
HIV	911 (9.8%)	683 (11.9%)	228 (6.4%)
TB current	112 (1.2%)	88 (1.5%)	24 (0.7%)
TB previously	643 (6.9%)	457 (8.0%)	186 (5.2%)
Asthma or COPD	1221 (13.1%)	951 (16.6%)	270 (7.5%)
Hypertension	5161 (55.5%)	4005 (70.0%)	1156 (32.3%)
CKD	852 (9.2%)	695 (12.1%)	157 (4.4%)
Pregnant	85 (0.9%)	58 (1.0%)	27 (0.8%)
Hospital admission	4181 (44.9%)	2474 (43.2%)	1707 (47.6%)
Admitted to ICU	373 (4.0%)	152 (2.7%)	221 (6.2%)
Ventilated	52 (0.6%)	15 (0.3%)	37 (1.0%)
New diabetes	1053 (11.3%)	229 (4.0%)	824 (23.0%)
Diabetes	9305 (100.0%)	5721 (100.0%)	3584 (100.0%)

HIV, Human immunodeficiency virus; TB, Tuberculosis; COPD, Chronic obstructive pulmonary disease; CKD, chronic kidney disease; ICU, intensive care unit

Supplementary Table 4.2. Counts of diabetes medication recorded as dispensed to public health sector patients with COVID-19 who had a diabetes comorbidity who accessed facilities with a computer-based pharmacy system. The results have been grouped by COVID-19 status (Recovered or Died).

	ALL *N=5708*	**Died** *N=928*	**Recovered** *N=4780*
Metformin	4084 (71.5%)	593 (63.9%)	3491 (73.0%)
Insulin	2073 (36.3%)	412 (44.4%)	1661 (34.7%)
Sulphonylurea	2110 (37.0%)	301 (32.4%)	1809 (37.8%)
Hydrochlorothiazide	2756 (48.3%)	400 (43.1%)	2356 (49.3%)
ACE inhibitor	2987 (52.3%)	486 (52.4%)	2501 (52.3%)
ARB	643 (11.3%)	150 (16.2%)	493 (10.3%)
Aspirin	1728 (30.3%)	373 (40.2%)	1355 (28.3%)
Statin	3925 (68.8%)	685 (73.8%)	3240 (67.8%)
Beta blocker	1363 (23.9%)	288 (31.0%)	1075 (22.5%)
Steroids	261 (4.6%)	54 (5.8%)	207 (4.3%)
Anti-retroviral therapy	583 (10.2%)	101 (10.9%)	482 (10.1%)
TB drugs	68 (1.2%)	17 (1.8%)	51 (1.1%)

ACE, Angiotensin converting enzyme; ARB, Angiotensin Receptor Blockers

Supplementary Table 4.3. Summary statistics of laboratory tests done on public health sector patients with COVID-19 and a diabetes comorbidity who were admitted into hospital. These results are the median (IQR) and proportions for tests done 2 days before and up to 5 days after the hospital admission for COVID-19. The results have been grouped by COVID-19 status (Recovered or Died).

	Died	**Recovered**	**N**
	N=1129	*N=2535*	
Creatinine	109.0 [76.0;172.0]	78.0 [60.0;106.0]	3459
eGFR	52.5 [29.0;60.6]	60.6 [52.0;60.6]	3450
eGFR:			3664
Stage 2 CKD	464 (41.1%)	1636 (64.5%)	
Stage 3A CKD	152 (13.5%)	297 (11.7%)	
Stage 3B CKD	191 (16.9%)	218 (8.6%)	
Stage 4 CKD	135 (12.0%)	142 (5.6%)	
Stage 5 CKD	136 (12.0%)	79 (3.1%)	
Unknown	51 (4.5%)	163 (6.4%)	
White Cell Count	9.9 [7.3;13.5]	8.3 [6.3;11.2]	3375
Lymphocyte count	1.1 [0.8;1.6]	1.4 [1.0;1.9]	2261
Ferritin	1084.0 [668.0;1749.0]	552.0 [299.0;1152.0]	390
Sodium	135.0 [131.0;139.0]	134.0 [131.0;137.0]	3119
D-Dimer	1.1 [0.5;4.1]	0.5 [0.3;1.0]	1410
D-Dimer			3664
high	472 (41.8%)	735 (29.0%)	
normal	37 (3.3%)	166 (6.5%)	
Unknown	620 (54.9%)	1634 (64.5%)	
C-Reactive Protein	176.0 [102.8;277.2]	113.0 [57.0;199.0]	2188
HbA1c	9.0 [7.1;11.6]	10.0 [7.3;12.6]	1761
HbA1c:			3664
<7%	119 (10.5%)	248 (9.8%)	
>9%	280 (24.8%)	731 (28.8%)	
7-9%	146 (12.9%)	237 (9.3%)	
Unknown	584 (51.7%)	1319 (52.0%)	

Supplementary Table 5.1: Characteristics of the study population, and stratified by HIV status

	Whole study population *N=10541*	**HIV negative** *N=8969*	**HIV positive** *N=1572*	**P value**
Sex: Female	7053 (67.0%)	5999 (67.0%)	1054 (67.1%)	0.947
Diabetes Ascertainment Age (Years):	52.0 [44.0;59.0]	53.0 [45.0;61.0]	46.0 [39.0;52.0]	<0.001
Diabetes Treatment Initiation Age (Years):	53.0 [45.0;60.0]	54.0 [46.0;62.0]	46.0 [40.0;53.0]	<0.001
Diabetes Treatment Formulation:				0.107
Metformin only	3525 (33.4%)	2982 (33.2%)	543 (34.5%)	
Metformin & Sulphonylurea	4417 (41.9%)	3796 (42.3%)	621 (39.5%)	
Metformin, Sulphonylurea & Insulin	2599 (24.7%)	2191 (24.4%)	408 (26.0%)	
Diabetes Treatment Initiation:				<0.001
At diabetes ascertainment	5828 (55.3%)	4936 (55.0%)	892 (56.7%)	
Within 1 year of ascertainment	2156 (20.5%)	1792 (20.0%)	364 (23.2%)	
More than 1 year after ascertainment	2557 (24.3%)	2241 (25.0%)	316 (20.1%)	
HIV Antiretroviral Treatment:	1202 (11.4%)	0 (0.0%)	1202 (76.5%)	0.000
Cluster:				<0.001
Adherent	1544 (14.6%)	1192 (13.3%)	352 (22.4%)	
Low adherence gradual decline	4656 (44.2%)	4005 (44.7%)	651 (41.4%)	
High adherence rapid decline	2716 (25.8%)	2380 (26.5%)	336 (21.4%)	
Low adherence gradual increase	1625 (15.4%)	1392 (15.5%)	233 (14.8%)	
Hypertension:	6517 (61.8%)	5713 (63.7%)	804 (51.1%)	<0.001
Tuberculosis:	1385 (13.1%)	850 (9.5%)	535 (34.0%)	<0.001

Supplementary Table 5.2: Median HbA1c values and proportion of study participants with HbA1c measures in the five years post diabetes ascertainment

	Adherent (D)	**Low adherence gradual decline (A)**	**High adherence rapid decline (B)**	**Low adherence gradual increase (C)**	
	N=1490	*N=3740*	*N=2591*	*N=1553*	**N**
Year one HbA1c (%)	9.40 [7.60;11.8]	9.00 [7.20;11.4]	9.30 [7.40;11.7]	9.10 [7.30;11.4]	7676
Participants with year one HbA1c	1320 (88.6%)	2881 (77.0%)	2244 (86.6%)	1231 (79.3%)	9374
Year two HbA1c	8.80 [7.20;11.1]	8.70 [7.10;11.1]	8.70 [7.10;11.0]	8.70 [7.20;11.0]	4430
Participants with year two HbA1c	806 (54.1%)	1431 (38.3%)	1319 (50.9%)	874 (56.3%)	9374
Year three HbA1c (%)	9.10 [7.40;11.3]	9.00 [7.30;11.4]	9.20 [7.30;11.4]	8.80 [7.30;11.1]	4001
Participants with year three HbA1c	679 (45.6%)	1451 (38.8%)	1170 (45.2%)	701 (45.1%)	9374
Year four HbA1c (%)	9.50 [7.50;11.5]	9.20 [7.30;11.4]	9.60 [7.70;11.9]	9.10 [7.40;11.3]	3765
Participants with year four HbA1c	585 (39.3%)	1420 (38.0%)	1115 (43.0%)	645 (41.5%)	9374
Year five HbA1c (%)	9.80 [7.80;11.7]	9.30 [7.50;11.3]	9.70 [7.80;11.7]	9.70 [7.60;11.4]	3250
Participants with year five HbA1c	445 (29.9%)	1304 (34.9%)	985 (38.0%)	516 (33.2%)	9374

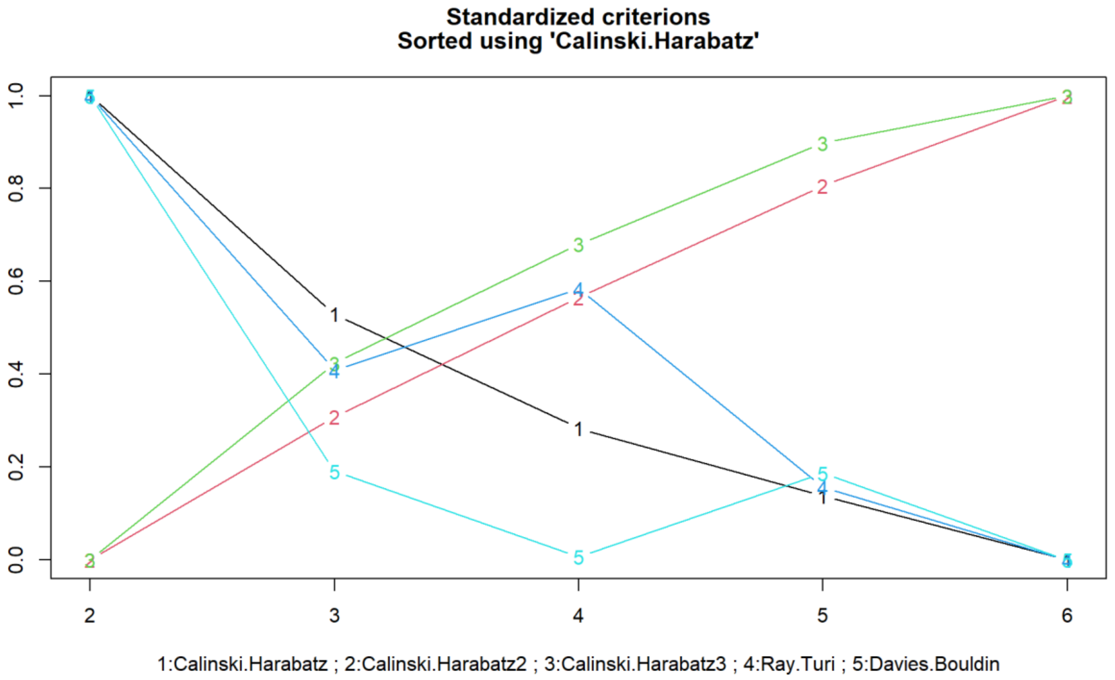
*N is the number of people in the study population who had an available HbA1c at the different time points.

Supplementary Table 5.3: Counts (%) and median (IQR) health facility encounters for study participants in the six months before starting diabetes treatment, in the two-year observation window (in four-month sliding windows) and in the 6 months after the two-year study observation window.

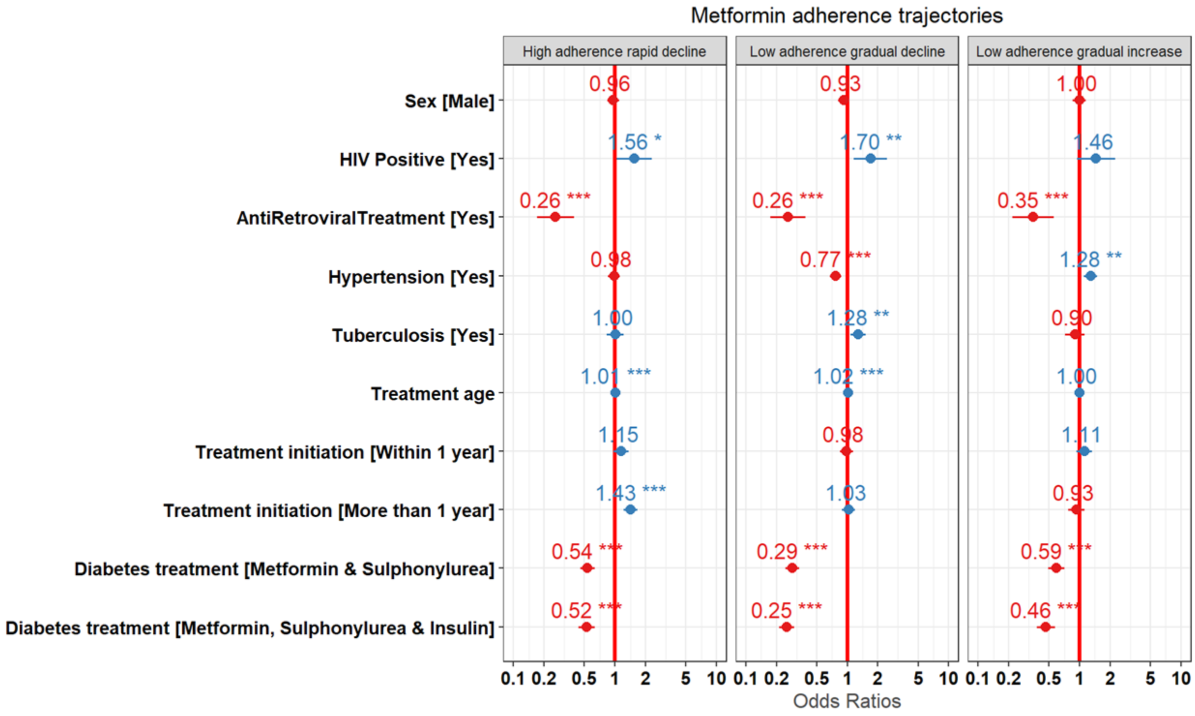
	Adherent (D) *N= 1544*	**Low adherence gradual decline (A)** *N= 4655*	**High adherence rapid decline (B)** *N= 2716*	**Low adherence gradual increase (C)** *N= 1625*
Six months before diabetes treatment start (median, IQR)	2.0 [1.0;4.0]	2.0 [0.0;4.0]	1.0 [0.0;3.0]	2.0 [1.0;4.0]
Six months before diabetes treatment start (%)	1172 (75.9%)	3309 (71.1%)	1848 (68.0%)	1248 (76.8%
Four months after diabetes treatment start (median, IQR)	5.0 [4.0;7.0]	3.0 [1.0;5.0]	4.0 [1.0;6.0]	4.0 [2.0;6.0]
Four months after diabetes treatment start (%)	1497 (97.0%)	4078 (87.6%)	2239 (82.4%)	1563 (96.2%)
Eight months after diabetes treatment start (median, IQR)	4.0 [2.0;5.0]	1.0 [0.0;3.0]	2.0 [1.0;4.0]	2.0 [1.0;4.0]
Eight months after diabetes treatment start (%)	1488 (96.4%)	2787 (59.9%)	2270 (83.6%)	1399 (86.1%)
Twelve months after diabetes treatment start (median, IQR)	4.0 [2.0;5.0]	1.0 [0.0;3.0]	2.0 [1.0;4.0]	2.0 [1.0;4.0]
Twelve months after diabetes treatment start (%)	1449 (93.8%)	2788 (59.9%)	2291 (84.4%)	1397 (86.0%)
Sixteen months after diabetes treatment start (median, IQR)	4.0 [2.0;5.0]	1.0 [0.0;3.0]	2.0 [1.0;4.0]	3.0 [1.0;4.0]
Sixteen months after diabetes treatment start (%)	1382 (89.5%)	2711 (58.2%)	2218 (81.7%)	1418 (87.3%)
Twenty months after diabetes treatment start (median, IQR)	3.0 [1.0;5.0]	1.0 [0.0;2.0]	2.0 [1.0;4.0]	3.0 [1.0;5.0]

	Adherent (D) *N= 1544*	Low adherence gradual decline (A) *N= 4655*	High adherence rapid decline (B) *N= 2716*	Low adherence gradual increase (C) *N= 1625*
Twenty months after diabetes treatment start (%)	1251 (81.0%)	2475 (53.2%)	2058 (75.8%)	1387 (85.4%)
Twenty-four months after diabetes treatment start (median, IQR)	3.0 [0.0;5.0]	1.0 [0.0;3.0]	2.0 [0.0;4.0]	3.0 [1.0;4.0]
Twenty-four months after diabetes treatment start (%)	1141 (73.9%)	2482 (53.3%)	1964 (72.3%)	1308 (80.5%)
Six months after study observation window (median, IQR)	4.0 [0.0;7.0]	2.0 [0.0;5.0]	4.0 [1.0;6.0]	4.0 [1.0;6.0]
Six months after study observation window (%)	1093 (70.8%)	2756 (59.2%)	2038 (75.0%)	1256 (77.3%)

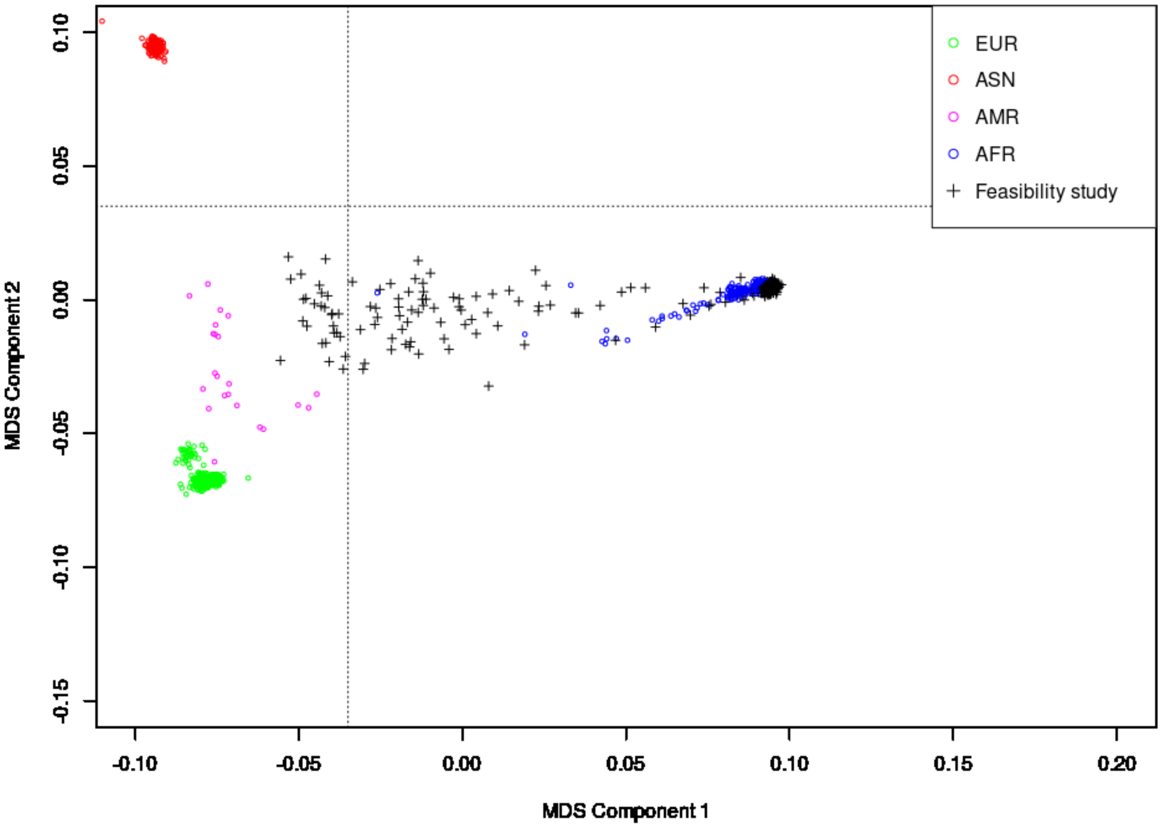
Supplementary figure 5.1



Supplementary figure 5.2



Supplementary figure 6.1



Supplementary file 6.1

PARTICIPANT INFORMATION

Who are we?

My name is Tsaone Tamuhla and I am a PhD student in the Faculty of Health Sciences at the University of Cape Town (UCT), and I am the main researcher in this study. I am working under the supervision of Associate Professor Nicki Tiffin and Professor Nicola Mulder at the University. We are also working with Dr Peter Raubenheimer and Dr Joel Dave, who care for patients with diabetes at Groote Schuur Hospital.

Why are we doing this study?

We want to study something called "genes". These "genes" are present in all of us and are the same in all parts of our bodies. "Genes" are sometimes also called DNA, which is the name of the material they are made from. Genes are responsible for why people in families are often more like each other, and different from other families. For example, some families are generally taller or shorter than others. This kind of information is passed from both the father and the mother to their children and on to their grandchildren, from one generation to the next. Some of these genes may prevent some people from getting certain illnesses. Other genes may be one of the reasons why some people get sick or have side effects from some medicines when others do not. We are still learning how genes might contribute to different diseases, and how they work together with our lifestyle and other factors - such as our environment or what we eat - to affect our health. We want to explore whether genes may affect type 2 diabetes in South African patients.

You may not get any benefit directly from this study, but we hope that the information we get about your genes and your health may benefit others who have diabetes and many different kinds of illnesses, in the future. You do not have to take part in this study, it is your choice if you want to take part, or not. If you do not want to take part, it will NOT affect the health care you receive at Groote Schuur Hospital.

We will also ask you if we can use your health and genetic information in other studies about diabetes, in the future; and if we can use your information in the future in other studies about different health conditions. You can choose whether you agree to this or not, and your decision will not affect your health care at Groote Schuur Hospital or your involvement in this study.

What results from this study will you receive?

We will not give you any individual results from the study of the samples you give us. This is because it will probably take a long time for this project to result in specific health information that is useful to patients.

If you agree for your information to be used in other studies in the future, it is possible that some new health information might be discovered in those studies. We will ask you if you would like to know any new information that might become available about your health.

At the end of the study, we will put our general findings from the study in some pamphlets and posters at the clinics where people have joined this study. There, you will be able to read how this study is contributing to our understanding of health and disease. When we describe the results of this study in this way, we will only show summary results or overall study results from the whole study, and there will be no information about the individual people who took part in the study.

What will we ask for?

We will ask you a few simple questions about your life and where you grew up. In order to better understand your health, we will ask you for permission to look at your health records that the Department of Health collects when you visit government health facilities like clinics or hospitals – such as any medical tests that have been done and any medication that you have been given at any government facility. We will also ask you to provide a swab or rinse of the inside of your mouth and we will use this to prepare a sample of your genes.

The sample will be stored at Stellenbosch University until the study is over. The information from your sample will be very securely stored at the University of Cape Town.

Sometimes researchers combine the information from everyone in the study and provide a summary of genetic information for the whole group. This does not provide information about each individual, but can provide information about the whole group all together. We will ask you whether you would agree for your information to be included in this kind of combined information.

How will we protect your information?

All your genes together make a special pattern in all of your body that only you have, and this is why no two people are exactly alike. Because each person has their own special pattern of genes, researchers are very careful to protect the genetic samples that are collected and the information from these samples, and these samples and information will only be used in the way you have agreed to.

To make sure that your privacy is protected in this study, we will make sure that your information is used for this research without your name, or your date of birth, or any other identifying information attached to it. This way, no one working on this study will know who the information or the genetic sample come from. We also want to make sure that your health information and sample are protected and safely stored, because there is always some small risk that the special pattern of genes from your sample could be used to work out who you are and see your health information if other people were to get hold of this information. We will be very careful in making sure all this personal information is very secure so that this could not happen.

The University of Cape Town will lock away any document with your name on it so that no-one can identify you from it. We will make sure all computers used for the study are kept securely and are protected by passwords.

What to do if you have questions or change your mind about being in the study.

If you have any questions, you can contact **The Human Research Ethics Committee Faculty of Health Sciences UCT** and speak to **Professor Marc Blockman on 021 406 6496**. If you change your mind and you no longer want your information or sample to be included in this study, or you have other questions you can also contact **Nicki Tiffin on 021 650 2506** with your questions, or to have your information removed from this study and to have your sample destroyed.

CONSENT FORM

Stick sample ID sticker here

Stick folder number sticker here

Name: ______________________________

Study ID number: ______________________________

Date of Birth: ______________________________

Height (cm): ______________________________

Weight (kg): ______________________________

What is your family origin? You can tick more than one box.

□ Xhosa □ Zulu □ Ndebele □ Swazi □ KhoiKhoi or San

□ Basotho □ Bapedi □ Tsonga □ Tswana □ Mixed Ancestry (Coloured)

□ Venda □ Afrikaans □ English □ Asian □ Other

Did you grow up in a town or a rural area? You can tick more than one box.

□ Town □ Rural

1. Do you agree for us to collect this saliva sample and your health information from the Department of Health for this study we have described about how genes might affect diabetes?

   □ YES □ NO

2. Do you agree for us to use your genetic sample together with your health information from the Department of Health for <u>other studies</u> in the future that want to study the effect of genes on diabetes?

   □ YES □ NO

3. Do you agree for us to use your genetic sample together with your health information for other studies in the future to study the effect of genes on <u>other health conditions</u> (not only diabetes) or biological functions?

□ **YES** □ **NO**

4. *(If 2. and/or 3. answered YES)* Sometimes, what we find from our research might include new information about your health. Would you like us to contact you again if we believe we have new information that may directly affect your health:

   If there is some kind of action or treatment that might be able to help you with the health issue?

   □ **YES** □ **NO**

   If there is NO kind of action or treatment that might be able to help you with the health issue?

   □ **YES** □ **NO**

5. Sometimes researchers combine the genetic information from everyone in the study and provide a summary of genetic data for the whole group. Do you agree for us to use your information when providing combined information about the whole research group (300 individuals in this study)?

   □ **YES** □ **NO**

6. Sometimes, what we find from a study like this might lead to new studies being done in the future. Can we contact you in the future to invite you to take part in other research studies like this one?

   □ **YES** □ **NO**

SIGNED: ______________________ **DATE:** ______________________

CONTACT DETAILS IF QUESTION NUMBER 4 and/or 6 IS ANSWERED 'YES':

____

Consent process undertaken by:

NAME: ______________________________

SIGNED: ______________________________ DATE: ___________________________________

Supplementary file 6.2

Ulwazi lomthathi-nxaxheba

Singoobani?

NdinguTsaone Tamuhla umfundi wePhD kwiSebe lezeMpilo kwiYunivesithi yaseKapa kwaye ndingumphandi ophambili koluphando. Ndisebenza phantsi komhloli uAssociate Professor Nicki Tiffin noProfessor Nicola Mulder kwiyunivesithi. Sisebenza noGqirha Peter Raubenheimer noGqirha Joel Dave, abanakekela abantu abanesifo seswekile kwisibhedlele iGroote Schuur.

Olu phando silwenzela ntoni?

Sifuna ukuphanda into ekuthiwa "ngamadlala emfuza" (genes). La madlala emfuza akhona kuthi sonke kwaye ayafana kuwo onke amalungu emizimba yethu. Ngamanye amaxesha amadlala emfuza kuthiwa yiDNA, eli ligama lento enziwe ngayo. Amadlala emfuza ngawo abangela ukuba abantu abakusapho olunye bafane, bangafani nabanye bezinye intsapho. Umzekelo, ezinye intsapho zinabantu abade ezinye abafutshane kunabanye. Lamadlala emfuza asuka kumama notata adlule aye kwabantwana nabazukulwana, nakwizizukulwana. Amanye amadlala emfuza enza ukuba abanye abantu bangazifumani izigulo ezithile. Amanye angunobangela okuba abantu bagule okanye babeneziphumo ezingalu lungelanga kumachiza bangakwazi ukusebenzisa amanye amayeza asetyenziswayo ngabanye abantu. Sisafunda ukuba amadlala emfuza enza igalelo elithini kwizigulo ezahlukeneyo kwaye zisebenza njani nendlela yethu yokuphila nezinye izinto – ezifana nezinto esizityayo okanye okusingqongileyo – ziyichaphazela njani impilo yethu. Sifuna ukufumanisa ukuba amadlala emfuza ayasichpahazela na isifo seswekile seType2 kwizigulane zaseMzantsi Afrika.

Unongazuzi ngqo wena koluphando, kodwa sinethemba lokuba ulwazi esilufumanayo ngamadlala emfuza akho kwilixa elizayo nempilo yakho lungayinzuzo kwabanye abanesifo seswekile nezinye izigulo ezahlukileyo. Awunyanzelekanga uthathe inxaxheba koluphando, sisigqibo sakho ukuba uyafuna okanye awufuni. Ukuba awufuni, oku akuzi kuchaphazela inkonzo zonyango olufumanayo kwisibhedlele iGroote Schuur.

Siza kucela imvume yakho ukuba sifuna ukusebenzisa iinkcukacha zakho zempilo kunye nezemfuza kolunye uphando lwesifo seswekile oluthe lwavela. Kwi xesha elizayo siza kucela imvume yakho kwakhona ukuba sifuna ukusebenzisa iinkcukacha zakho kuphando kolunye uphando lwezigulo oluthe lwavela. Ungakhetha ukuvuma okanye ukungavumi kwaye isigqibo sakho, nenxaxheba yakho koluphando, asizi kuchaphazela inkonzo zakho zonyango kwakho kwisibhedlele iGroote Schuur.

Uza kufumana iziphumo ezithini koluphando?

Asizi kukunika mntu iziphumo ezivele kwisampulu osinika yona. Le nto inje kuba kuza kuthatha ixesha elide ukuba olu phando lukhuphe iziphumo zempilo eziluncedo kwizigulane.

Ukuba uyavuma ukuba iinkcukacha zakho zisetyenziswe kolunye uphando, kungenzeka ukuba ulwazi olutsha lwezempilo lufamaneke. Siza kukubuza ukuba uyafuna ukwazi ulwazi olutsha oluvelayo ngempilo yakho.

Ekugqibeleni kophando, siza kufaka iziphumo zethu ngokubanzi kwiiphamflethi neepowusta kwiziko lempilo apho abantu bathathe inxaxheba koluphando. Apho uza kwazi ukufunda ngokuba olu phando lwenza ugalelo olungakanani ekuqondeni kwethu ngempilo nezifo nezigulo. Xa sichaza iziphumo zophando, siza kukhupha isishwankathelo seziphumo okanye iziphumo zizonke zophando; akuzi kubakho ziinkcukacha zabantu abathathe inxaxheba kuphando.

Siza kucela okanye sikubuze ntoni?

Siza kukubuza imibuzo embalwa, elula ngobomi bakho nokuba ukhulele phi. Ukuze sikwazi ukuqonda impilo yakho kakuhle, siza kucela imvume yakho ukuba sijonge iifayili zakho zempilo eziqokelelwa kwiSebe lezeMpilo xa usiya kwiiklinikhi nezibhedlele zikarhulumente – izinto ezifana namayeza owafumanayo novavanyo lwempilo olwenzayo kwisibhedlele okanye iklinikhi karhulumente. Siza kucela kwakhona ukuba usinike amathe akho ngokuthi siwathathe ngento apha kuwe emlonyeni okanye usele amanzi uwathufele entweni; la mathe siza kuwasebenzisa ukwenza isampulu yakho yamadlala emfuza.

Isampulu iza kugcinwa kwiYunivesithi iStellenbosch lude uphando luphele. Ulwazi olufunyenwe kwisampulu yakho luza kugcinwa ngokukhuselekileyo kwiYunivesithi yaseKapa.

Ngamanye amaxesha abaphandi bayaludibanisa ulwazi olufunyenwe kubathathi-nxaxheba ukuze benze isishwankathelo solwazi lwamadlala emfuza oluquka bonke abathathi-nxaxheba. Oku akukhuphi lwazi ngomntu ngamnye kodwa likhupha ulwazi ngeqela lonke lidibene. Siza kubuza ukuba uyavuma na ukuba ulwazi olufumaneke kuwe lungasetyenziswa ngoluhlobo.

Siza kuzikhusela njani iinkcukacha zakho?

Xa ewonke, amadlala emfuza akho enza ufuzo emzimbeni wakho ongowakho wedwa, yiyo le nto kungekho bantu abafana ncakasana. Ngenxa yokuba umntu ngamnye enomfuziselo wakhe wamadlala emfuza, abaphandi benza ngocoselelo ukuzikhusela iisampulu ezithathiweyo neenkcukaha zempilo eziphuma kwezisampulu, kwaye ezi sampulu neenkcukacha ziza kusetyenziswa ngale ndlela uyivumeleyo kuphela.

Ukuqinisekisa ukuba ukhuselekile koluphando, siza kuqinisekisa ukuba iinkcukacha zakho zisetyenziswa kolu phando ngaphandle kwegama lakho, umhla wakho wokuzalwa, naluphi na ulwazi olunokuveza ukuba ungubani na. Ngolu hlobo, abantu abasebenza kolu phando abazi kwazi ukuba nolwazi okanye isampulu yamadlala emfuza isuka kubani na. Sifuna nokuqinisekisa ukuba iinkcukacha zakho zempilo nesampulu yakho zikhuselekile kuba kukho umngcipheko omncincane wokuba umfuziselo wakho wamadlala emfuza ungasetyenziswa ukufumanisa ukuba ungubani, kubonwe iinkcukacha zakho zempilo. Siza kusebenza ngocoselelo ukuqinisekisa ukuba zonke iinkcukacha zakho eziyimfihlelo zikhuseleke ngeyona ndlela ukuze oku kungenzeki.

IYunivesithi yaseKapa iza kuwatshixela onke amaphepha anamagama akho ukuze kungabikho mntu ukwaziyo ukukuchaza ngawo. Siza kuqinisekisa ukuba zonke iikhompyutha ezisetyenziswa kolu phando zigcinwa kwindawo ekhuselekileyo kwaye zikhuselwe ngeepasiwedi.

Ungenza ngani ukuba unemibuzo okanye utshintsha ingqondo ngokuthatha inxaxheba.

Ukuba kunemibilizo onayo, ungafowunela **iHuman Research Ethics Committee Faculty of Health Sciences UCT** uthethe no**Profesa Marc Blocknam** ku**021 406 6496**. Ukuba utshintsha ingqondo yakho kwaye awusafuni ukuthatha inxaxheba kolu phando, okanye uneminye imibuzo ungafowunela u**Nicki Tiffin** ku**021 650 2506,** ungacela nokuba iinkcukacha zakho zikhutshwe kolu phando kwaye nesampulu yakho.

IFOMU YEMVUME

Stick sample ID sticker here

Stick folder number sticker here

Igama: ____________________

Inombolo yeStudy ID: ____________________

Umhla wokuzalwa: ____________________

Ubude (cm): ____________________

Ubunzima (kg): ____________________

Ithini imvelaphi yosapho lwakho? Ungakhetha iimpendulo eziliqela.

□ abaXhosa
□ abeSuthu
□ abaVenda
□ abaZulu
□ abaPedi
□ amaBhulu
□ abaNdebele
□ abaTsonga
□ abeLungu
□ abaSwazi
□ abaTswana
□ abaseAsiya
□ abaKhoiKhoi/ San
□ abantu beBala
□ Abanye

Ukhulele edolophini okanye elalini? Ungakhetha iimpendulo eziliqela.

□ Edolophini □ Ezilalini

7. Uyavuma ukuba sithathe le sampulu yamathe akho kwaye neenkcukacha zakho zempilo kwiSebe lezeMpilo kolu phando siluchazileyo lokuba amadlala emfuza anegalelo elinjani kwisifo seswekile?

□ EWE □ HAYI

8. Uyavuma ukuba sisebenzise isampulu yakho kwaye neenkcukacha zakho ezivela kwiSebe lezeMpilo kolunye uphando oluthe lwavela olufuna ukuphanda iimpembelelo zamadlala mfuza kwisifo seswekile?

□ EWE □ HAYI

9. Uyavuma ukuba sisebenzise isampulu yakho neenkcukacha zakho ezivela kwiSebe lezeMpilo kolunye uphando oluthe lwavela olujonga igalelo lamadlala emfuza kwezinye izigulo (ngaphandle kwesifo seswekile) okanye ukusebenza kwendalo yomzimba?

□ **EWE** □ **HAYI**

10. *(Ukuba uphendule "EWE" ku2 okanye ku3)* Ngamanye amaxesha sifumanisa ukuba uphando lwethu luveza ulwazi olutsha ngempilo yakho. Ungathanda ukuba sikufowunele kwakhona ukuba sikholelwa ukuba sinolwazi olutsha olunokuchaphazela impilo yakho:

Ukuba kukho unyango oluthile okanye into onokuyenza enokunceda isigulo sakho?

□ **EWE** □ **HAYI**

Ukuba akukho nyango olukhoyo okanye nto onokuyenza ukunceda isigulo sakho?

□ **EWE** □ **HAYI**

11. Ngamanye amaxesha abaphandi bayaludibanisa ulwazi olufunyenwe kwabathathi-nxaxheba bophando ukuze benze isishwankathelo lolwazi lwamadlala emfuza oluquka bonke abathathi-nxaxheba. Uyavuma ukuba sisebenzise olufunyenwe ngawe ulwazi xa sisenza isishwankathelo seqela lonke (kukho abathathi-nxaxheba abayi-300)?

□ **EWE** □ **HAYI**

12. Ngamanye amaxesha sifumanisa ukuba uphando olunje lungakhokhelela kuphando olungolunye. Singakufowunela sikumeme uthathe inxaxheba kolunye uphando olunje ngolu?

□ **EWE** □ **HAYI**

ISAYINWE: ______________________________ **UMHLA:** ___________________________________

IINKCUKACHA ZAKHO ZONXIBELELWANO UKUBA UPHENDULE 'EWE' KUMBUZO 4 OKANYE UMBUZO 6.

____

Inkqubo yemvume eyenziwa ngu (consent process undertaken by):

IGAMA: ______________________________

ISAYINIWE: ______________________________ UMHLA: ___________________________________

Supplementary Table 6.1. Allelic counts of known COVID-19 variants in the feasibility study population.

SNP	Location#	Allelic counts in cases (n= 63)	Allelic counts in controls (n= 280)	Odd ratio	P-value
rs9427097	Chr 1, 154568683	0	0	NA	NA
rs1801274	Chr 1, 161479745	58	282	0.8408	0.3803
rs3766539	Chr 1, 203193261	0	0	NA	NA
rs1800896	Chr 1206946897	0	0	NA	NA
rs360102	Chr 1, 226067862	46	168	1.335	0.1617
rs7595310	Chr 2, 168810137	50	153	1.75	0.006024
rs10490770	Chr 3, 45864732	0	0	NA	NA
rs2282679	Chr 4, 72608383	0	0	NA	NA
rs1173773	Chr 5, 32750983	0	0	NA	NA
rs17448496	Chr 5, 146015615	16	44	1.706	0.08222
rs155788	Chr 5, 179260528	0	0	NA	NA
rs17142392	Chr 6, 6626983	48	217	0.9727	0.8915
rs1799945	Chr 6, 26091179	6	10	2.75	0.04551
rs3131294	Chr 6, 32180146	1	7	0.632	0.6664
rs2069837	Chr 7, 22768027	19	80	1.065	0.8188
rs657152	Chr 9, 136139265	59	248	1.108	0.6045
rs579459	Chr 9, 136154168	0	0	NA	NA
rs1800450	Chr 10, 54531235	0	0	NA	NA
rs2957707	Chr 11, 10377258	0	0	NA	NA
rs2923084	Chr 11, 10388782	51	234	0.9474	0.7875
rs10774671	Chr 12, 113357193	54	289	0.7033	0.07593
rs10735079	Chr 12, 113380008	40	128	1.555	0.04015
rs7318817	Chr 13, 28617708	0	0	NA	NA
rs1048943	Chr 15, 75012985	0	0	NA	NA
rs13334749	Chr 16, 4952194	0	0	NA	NA

SNP	Location#	Allelic counts in cases (n= 63)	Allelic counts in controls (n= 280)	Odd ratio	P-value
rs1024611	Chr 17, 32579788	28	86	1.575	0.06143
rs1042542	Chr 17, 76221428	0	0	NA	NA
rs4800182	Chr 18, 21812972	39	209	0.7486	0.1703
rs12979860	Chr 19, 39738787	0	0	NA	NA
rs1006111	Chr 19, 52717232	0	0	NA	NA
rs12329760	Chr 21, 42852497	0	0	NA	NA

#Location is the chromosome number and base pair of the SNP

Supplementary Table 6.2. Allelic counts of known type 2 diabetes variants in the feasibility study population.

SNP	Location#	Allelic count in cases (n= 93)	Allelic count in controls (n= 250)	Odds ratio	P-value
rs880315	Chr 1, 10796866	0	0	NA	NA
rs1801133	Chr 1, 11856378	24	31	2.26	0.0037
rs3480	Chr 1, 33328165	100	235	1.311	0.1152
rs2495477	Chr 1, 55518467	83	160	1.712	0.002118
rs12742393	Chr 1, 162224586	61	82	2.488	2.606e-06
rs1260326	Chr 2, 27730940	0	0	NA	NA
rs2943641	Chr 2, 227093745	0	0	NA	NA
rs11708067	Chr 3, 123065778	24	66	0.9742	0.9185
rs6444082	Chr 3, 185536223	0	0	NA	NA
rs3887925	Chr 3, 186665645	0	0	NA	NA
rs2282679	Chr 4, 72608383	0	0	NA	NA
rs1799883	Chr 4, 120241902	0	0	NA	NA
rs10033601	Chr 4, 153252061	103	214	1.658	0.003313
rs2255137	Chr 4, 153309538	0	0	NA	NA
rs17080093	Chr 6, 150997440	0	0	NA	NA
rs622342	Chr 6, 160572866	44	97	1.287	0.2201
rs16147	Chr 7, 24323410	0	0	NA	NA
rs3757840	Chr 7, 44231216	0	0	NA	NA
rs3808607	Chr 8,	0	0	NA	NA

SNP	Location$^{\#}$	Allelic count in cases (n= 93)	Allelic count in controls (n= 250)	Odds ratio	P-value
	59412924				
rs13266634	Chr 8, 118184783	0	0	NA	NA
rs2466293	Chr 8, 118185938	52	65	2.597	3.653e-06
rs7914287	Chr 10, 69350563	0	0	NA	NA
rs7079157	Chr 10, 71119208	0	0	NA	NA
rs2305198	Chr 10, 71128875	0	0	NA	NA
rs2637248	Chr 10, 78273721	57	152	1.012	0.9505
rs7903146	Chr 10, 114758349	0	0	NA	NA
rs12255372	Chr 10, 114808902	0	0	NA	NA
rs1001179	Chr 11, 34460231	0	0	NA	NA
rs1695	Chr 11, 67352689	76	263	0.6226	0.006257
rs2846707	Chr 11, 102576358	0	0	NA	NA
rs1044471	Chr 12, 1896956	0	0	NA	NA
rs1234032	Chr 12, 42354629	0	0	NA	NA
rs9581943	Chr 13, 28493997	49	40	4.113	2.064e-10
rs2470893	Chr 15, 75019449	0	0	NA	NA
rs16948048	Chr 17, 47440466	51	130	1.075	0.7077
rs8089787	Chr 18, 19406601	0	0	NA	NA
rs17782313	Chr 18, 57851097	0	0	NA	NA
rs1799817	Chr 19, 7125297	51	108	1.371	0.1083
rs2059806	Chr 19, 7166376	0	0	NA	NA
rs895819	Chr 19, 13947292	0	0	NA	NA
rs13037490	Chr 20, 23583725	0	0	NA	NA
rs1042531	Chr 20, 56140980	0	0	NA	NA
rs2825115	Chr 21, 20156686	0	0	NA	NA

#Location is the chromosome number and base pair of the SNP