Dissertation Type: Research

DEPARTMENT OF COMPUTER SCIENCE

The Challenges and Opportunities for Using Computational Notebooks in Co-Design: A Case Study of T1 Diabetes Machine Learning Co-Design

Jacob Hughes

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Bachelor of Science in the Faculty of Engineering.

Friday, 13 May 2022

Word Count: 15438

Abstract

This dissertation is a reflective account of using computational notebooks in co-design.

I performed a deductive thematic analysis of interviews with participants, who took part in a research study co-designing machine learning systems for type 1 diabetes care, using Jupyter Notebook.

The research objectives of this project are:

- Identify opportunities and challenges of using Jupyter Notebook as a collaborative tool for co-design
- Discuss changes to the Jupyter Notebook used in the COTADS project in accordance with the findings
- Discuss future work for the general use of Jupyter Notebook as a collaborative tool for co-design

Dedication and Acknowledgements

This project would not have been possible for me to complete without the extensive support I have received from a variety of people, which I am incredibly grateful for.

Firstly, I'd like to thank my supervisor, Dr Aisling O'Kane, for her continued support, guidance and patience which allowed me to pursue a project I was truly passionate in, despite setbacks and delays.

Additionally, I'd like to thank Dr Amid Ayobi for his role in the creation of COTADS workshop material, and for his time and support for me and my goals.

Thank you to Sam James for your consistent support in COTADS workshops and appearance in project meetings, and for your transcript text extraction script, which certainly saved me hours of time!

I'd also like to extend thanks to all the participants of the COTADS workshops.

Finally, I'd like to thank Niamh MacGloin for her relentless support, motivation, and faith in me.

Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Taught Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, this work is my own work. Work done in collaboration with, or with the assistance of others, is indicated as such. I have identified all material in this dissertation which is not my own work through appropriate referencing and acknowledgement. Where I have quoted or otherwise incorporated material which is the work of others, I have included the source in the references. Any views expressed in the dissertation, other than referenced material, are those of the author.

Jacob Hughes, Friday, 06 May 2022

Table of Contents

List of Tables

Ethics Statement

An ethics application for this project was reviewed and approved by the faculty research ethics committee as application 9164.

Supporting Technologies

- The research workshops were centred around the use of **Jupyter Notebook**, a web-based Python environment. When interviewing participants, they were asked to open the COTADS Notebook for reference during our discussion. I didn't use Jupyter Notebook after this stage.
- I used **Microsoft Teams** to conduct video-call interviews with participants, which also involved them sharing their screen. These video calls were recorded live with Teams' built-in recording feature.
- Recordings from Teams interviews were then securely uploaded to **Microsoft Stream**, a web service that stores the videos in an encrypted manner. The website can also be used to generate a closed-caption transcript of the video, which could then be downloaded and edited.
- I used a **bespoke Python script**, written by my colleague **Sam James**. The script takes the raw '.vtt' transcript file that MS Stream produces and removes the metadata contained in the file, such as timestamps and tokens, and converts it into a .txt file containing only the text transcript of the interview.
- I then used **Microsoft Word** to curate the text into legible, accurate transcriptions that could then be used for thematic analysis.
- I used the qualitative data analysis software, **NVivo**, to perform inductive thematic analysis on the transcripts of my exit interviews with workshop participants. NVivo allowed me to label quotes as codes, and then distil those codes into themes.

Notions and Acronyms

Listed below are some common acronyms that appear within the dissertation.

AI	:	Artificial Intelligence
CGM	:	Continuous Glucose Monitoring, *a device used by some diabetics that regularly measures the blood glucose levels of the user*
CLI	:	Clinician
COTADS	:	Co-Designing Trustworthy Autonomous Diabetes Systems
FC	:	Family Carer
HCI	:	Human Computer Interaction
JN	:	Jupyter Notebook
ML	:	Machine Learning
T1D	:	Type 1 Diabetes
UHS	:	University Hospital Southampton
UoB	:	University of Bristol
UoS	:	University of Southampton
YA	:	Young Adult

1. Introduction

In recent history, the scope of application of artificial intelligence and machine learning has had significant growth in many domains, including healthcare. While there are a diverse range of uses of AI and ML in healthcare, there are controversies surrounding its use, such as AI explainability [1], bias [2] and ethics [11]. The opportunity exists to use participatory methods, such as co-design[6], to take a human-centred approach when designing AI and ML healthcare systems. In this thesis, I investigate the methodology of using computational notebooks, a data science tool, as a tool for co-design of ML for healthcare, specifically Type 1 Diabetes.

A specific and well-known computational notebook is Jupyter Notebook. It is used for a wide variety of data science activities, but the aim of this dissertation is to investigate and evaluate the effectiveness of using Jupyter Notebook (JN) as the centrepiece for collaborative design workshops that involve participants from diverse professional backgrounds with varying (in many cases, zero) experience in computer science and Jupyter Notebook. While the use of JN as a collaborative tool between data scientists is well-established [3], the successful use of JN with non-expert users as a co-design tool is scarcely reported.

This project builds on the work of Ayobi et al [13], who have identified that co-design has 'social, technological, and strategic benefits' but also 'organisational, translational, and pragmatic challenges'. Expanding on their findings, this project takes a deeper look at the benefits and limitations of using Jupyter Notebook as a tool for co-design by analysing the methodology of its use in our project. The use-case of JN as a tool for co-design with non-experts has not been explored before in academic publishing. These novel findings could become the basis for a new use-case of JN.

I used deductive thematic analysis techniques to analyse interviews with research participants.

The research objectives of this project are:

- Identify opportunities and challenges of using Jupyter Notebook as a collaborative tool for co-design
- Discuss changes to the Jupyter Notebook used in the COTADS project in accordance with the findings
- Discuss future work for the general use of Jupyter Notebook as a collaborative tool for co-design

2. Related Work

2.1 History of Co-design

The history of co-design is rooted in 1970s Scandinavia, when a project with the Norwegian Iron and Metal Workers Union used revolutionary techniques of developing their computer systems with the end-users [4]. The project's successful approach reflected the value of incorporating the user in system development, and early adopters looked to pull on this thread. They found that the discussing the users' needs and experience of using the system was beneficial to the development of the system. This led to the inception of the UTOPIA project, who's research objectives were to 'contribute to the development of methods for involving end users in all phases of design and development of IT' [5]. The success of the UTOPIA project established the methodology of participatory design, incorporating end users and developers when designing software, and the benefits of using this methodology.

The modern definition of co-design is 'a practice where people collaborate or connect their knowledge, skills and resources in order to carry out a design task' [6]. This is achieved by bringing people, often from a wide variety of professional backgrounds, together into a team. In the case of systems design, the team should comprise of professional experts, researchers, and users of the system.

2.2 Co-design in healthcare

There is a call for co-design in healthcare, as vast amounts of clinical research findings are discovered to be unusable, resulting in the loss of time, money, and resources [7]. One of the reasons why clinical research can be unhelpful is due to its lack of patient-centredness – the research does not reflect the patient's priorities. By co-designing with medical researchers and patients, researchers can glean more insight into the patient's needs and steer research based on their findings, especially when patients are engaged towards the start of research [8].

Researchers have found that preliminary co-design of healthcare research leads "research topics, research questions and design of materials to be more applicable and acceptable to research end-users" [9].

When co-designing for personal health, an inevitable outcome is that co-researchers are afflicted or affected by the subject being researched. This means that certain sensitivities, precautions, and care must be taken when working with patients, but with "adequate infrastructure to accommodate and empower the co-researchers" [10], beneficial research can still be conducted.

2.3 Codesign of machine learning for healthcare

The call for co-designing ML in healthcare is growing, as healthcare technology becomes more ambitious. Co-designing medical technology also needs to be rigorously monitored to ensure that the technologies being developed are ethically sound, with considerations towards diversity, representation, access to technology and, of course, patient-centredness. [11]

Further issues with the complexity of ML concepts makes explaining ML to a lay audience difficult to approach, but through teaching methods such as data visualisation and analogies [12], there are ways to explain ML to a lay audience and achieve positive understanding.

Ayobi et al's paper, 'Co-Designing Personal Health? Multidisciplinary Benefits and Challenges in Informing Diabetes Self-Care Technologies' [13] is seminal to this dissertation. Ayobi et al outlined the methodology that is used in this dissertation; using qualitative analysis of interviews with participants of co-design workshops to identify benefits and challenges of co-design. While Ayobi et al investigated the methodology of co-designing ML solutions for T1D care, this dissertation looks one stage deeper,

and investigates the methodology of using Jupyter Notebook as a tool for co-designing ML solutions for T1D care.

2.4 Towards collaborative and interactive computational notebooks

A computational notebook is ‘an interactive web tool that can execute code, and display text and multimedia resources in a single document’ [14]. Computational notebooks are popular with data scientists because they afford the exploration of data, by providing an environment for code execution and visualisation. One popular computational notebook is Jupyter Notebook.

Jupyter Notebook “supports over 40 programming languages”, most importantly Python, and “can produce rich, interactive output: HTML, images, videos, LaTeX, and custom MIME types.” [15]

While Jupyter Notebook is used as a data science tool, due to its customisable and stylish interactive output, there is an opportunity to explore the use JN as an information presentation tool with interactive elements, that can help connect end users and designers.

The viability of the notebook paradigm as a collaborative tool has already been explored in industry, such as with the design suite ‘CoCalc’ [16] which boasts a collaborative whiteboard system that can host Jupyter Notebook code cells.

3. Methodology

3.1 Project Overview

The Co-Designing Trustworthy Autonomous Diabetes Systems (COTADS) project is a collaborative project between HCI researchers at the University of Bristol, ML researchers at the University of Southampton, and clinicians at University Hospital Southampton. The directive of the COTADS project is to design clinically trustworthy artificial intelligence algorithms that assist with type 1 diabetes (T1D) management.

Before I joined the project, the COTADS team had been conducting co-design sessions amongst themselves to develop and pilot the first version of the notebook. I conducted individual hour-long exit interviews regarding these workshops with the members of the team, which produced empirical feedback for the design of the notebook content, and methodological feedback for the aesthetic design and usage of the notebook.

As part of the project, the research team at UoB invited members of the public to attend a series of co-design workshops to explore the 'anticipated benefits and unintended consequences' of diabetes predictive care technology. The secondary directive of the workshops was to learn more about people's experiences of working in a collaborative, online environment.

The recruitment criteria for participants was:

- young adults aged 18-25 who live with type 1 diabetes
- or adults aged 18+ who care for a minor who is living with type 1 diabetes

According to JDRF, a UK-based type 1 diabetes charity, there are approximately 400,000 people living with T1D in the UK [17]. Due to the scarcity of qualifying applicants, the participants were recruited through 'snowball sampling' methods [18], which consisted of posting adverts on personal social media, diabetes-related social media groups, and forums such as the JDRF newsletter. Viewers of the advert were asked to pass on the advert to their friends and family. Through this method, ten participants were recruited, six young adults living with T1D, and four family carers of children living with T1D.

The workshops did not require participants to have any prior knowledge, besides lived experience of diabetes care.

3.2 Overview of the Co-Design Workshops

The ten participants were invited to participate in a total of five online workshops, performed via the video calling app Zoom, and supported with the use of the notebook. Participants were compensated for their time with up to £90 value of Amazon vouchers.

The notebook used in the workshops has been published open source [19].

Each workshop lasted one hour, and workshops were conducted in two separate groups, one with family carers and one with young adults. In anticipation of potential friction between the two participant groups, there were no sessions where young adults and family carers were discussing workshop activities in the same group. In the young adult workshops, when there was 6 participants present, the workshops were split into breakout rooms, with 3 participants in each room. If there was less than 6 participants then the workshop remained as one group.

In the first workshop for each group, the notebook was introduced and the workshop host used screen sharing to demonstrate the usage of the Jupyter Notebook. At the start of subsequent workshops, participants were reminded how to interact with the notebook, particularly the notion of executing a code cell.

During the workshops, as workshop leaders, HCI researchers presented a section of the notebook to the participants. Participants were given some time to interact with the activity, and then asked to discuss their thoughts on the content. Participants were encouraged to steer the discussion themselves, and workshop leaders would ask empirical questions about the notebook content in order to continue the conversation. Workshop leaders were also present as technical support for the participants if they needed help, and participants were encouraged to ask for help whenever they needed it.

Participants were not asked to work on the notebooks outside of the workshop sessions or perform any extra research. While prior knowledge of ML or JN was not a requisite for the workshop, participants YA2 and YA5 had extensive prior knowledge of both topics as they were both university students studying computer science and data science. Other participants, such as FC2 and YA4 reported that they had some limited prior experience of coding.

In the fourth workshop, we used a shared slide deck as a de facto whiteboard, and asked participants to share ideas as ‘post-it notes’ by typing their thoughts into text boxes on the slides. In the fifth and final workshop, each participant individually screen shared their notebook to the group and presented the feature importance graph that they had created.

Table 3.1: Table of Young Adults workshop schedule

#	Workshop date	Time	Notebook section(s) discussed	# of participants	Topics covered	Research methods
1	Mon 21 Feb	19:00-20:00	Section 1	6	Health and wellbeing data, Blood glucose level data	Group discussion, notebook interaction
2	Sun 27 Feb	19:00-20:00	Section 2	6	T1DExchange dataset, data and representation	Group discussion, notebook interaction
3	Wed 2 March	19:00-20:00	Section 3, 4	5	ML Model explanation, Risk score slider	Group discussion, notebook interaction
4	Sun 13 March	19:00-20:00	Section 5	4	Feature importance graph example	Group discussion, notebook interaction, slide deck and post-it notes
5	Sun 20 March	19:00-20:00	Section 6	6	Designing own feature importance graph	Group discussion, notebook interaction

Table 3.2: Table of Family Carer workshop schedule

#	Workshop date	Time	Notebook section(s) discussed	# of participants	Topics covered	Research methods
1	Wed 23 Feb	19:00-20:00	Section 1	3	Health and wellbeing data, Blood glucose level data	Group discussion, notebook interaction
2	Sun 6 March	19:00-20:00	Section 2	4	T1DExchange dataset, data and representation	Group discussion, notebook interaction
3	Tue 8 March	19:00-20:00	Section 3, 4	1*	ML Model explanation, Risk score slider	Group discussion, notebook interaction
4	Sun 13 March	19:00-20:00	Section 3,4,5	4	Risk score slider, Feature importance graph example	Group discussion, notebook interaction, slide deck and post-it notes
5	Sun 20 March	19:00-20:00	Section 6	4	Designing own feature importance graph	Group discussion, notebook interaction

* Since only one participant attended this session, this session was ran individually with other participants and a recap was performed in the following session.

Table 3.3: Table of workshop participants

#	ID	Role	Developer of Notebook	Prior experience of coding	Prior experience with notebook
1	HCI1	HCI researcher, Workshop Host	Yes	N/A	N/A
2	HCI2	HCI researcher, Workshop Host	Yes	N/A	N/A
3	HCI3	HCI researcher, Workshop Host	Yes	N/A	N/A
4	HCI4	HCI researcher, Workshop Host	No	N/A	N/A
5	HCI5	HCI researcher, Workshop Host	No	N/A	N/A
6	ML1	ML researcher, Workshop Guest	Yes	N/A	N/A
7	ML2	ML researcher, Workshop Guest	Yes	N/A	N/A
8	ML3	ML researcher	Yes	N/A	N/A
9	CLI1	Clinician	Yes	Some	None
10	CLI2	Clinician	Yes	None	None
11	CLI3	Clinician	Yes	None	None
12	YA1	Young adult with T1D	No	None	None
13	YA2	Young adult with T1D	No	Expert	Expert
14	YA3	Young adult with T1D	No	None	None
15	YA4	Young adult with T1D	No	Some	None
16	YA5	Young adult with T1D	No	Expert	Expert
17	YA6	Young adult with T1D	No	None	None
18	FC1	Family carer of child with T1D	No	None	None
19	FC2	Family carer of child with T1D	No	Some	None
20	FC3	Family carer of child with T1D	No	None	None
21	FC4	Family carer of child with T1D	No	None	None

3.3 Interview Process

After the workshops finished, I interviewed each participant individually. These interviews were conducted over Microsoft Teams and took between 1-1.5 hours. These interviews were video recorded and screen sharing was used. In the findings, observations of user interaction with the notebook while screen sharing were used and enhanced the data corpus from the interviews.

The interview questions were based on the questions asked in the exit interviews conducted with the COTADS team at the start of the term, with some revisions and additional questions to glean more feedback about the use of Jupyter Notebook in co-design.

The following is an overview of the interview process:

- Participants were informed that the purpose of the interview was to identify benefits and limitations of using computational notebooks for co-design. They were asked to not hold back any opinions and were encouraged to be critical.
- Participants were reminded of the ethics protocol, and reminded that they will be able to review and edit any quotes prior to publication, and that they would not be identified as a named individual in publication.
- Participants were asked permission to record the audio and screen of the interview.

The first section of questioning was general questions about the use of computational notebooks. This involved questions such as:

- In general, what are your thoughts on how we have used the computational notebook?
- How would describe your first impressions of using the computational notebook as part of this project?
- How would you compare your experience using the computational notebook with other tools, for example, MS PowerPoint?
- Has your educational / professional background affected your capacity to interact with this project / computational notebook?

The next section went through each workshop session individually, and asked questions about the activities that occurred in that session. The questions asked were specific to each session, but generally sought the same information for each workshop:

- What was your experience of interacting with this activity?
- Is there anything you would change about this activity?
- Were there any challenges with this activity?
- Did the activity meet your expectations?
- What did you like / dislike about this activity?

The next section investigated how well the Notebook had explained machine learning concepts:

- How would you describe your current understanding of the ML model?
- How did the ML model explanations, such as the feature importance graph and the interactive risk score, support you in understanding the ML model?

The next section addressed co-design and how participants felt about working together:

- How did you find working with people from other disciplines?
- How would you describe the role of the notebook when working with other people?
- Can you think of any way we could improve the collaboration?
- What are your thoughts on having mixed groups (family carers and young adults together)?

The concluding section was overall questions:

- How would you describe your experience of taking part in the five workshops we've conducted?
- What was your favourite part?
- How do you feel about the workshop outcomes considering your expectations?

- Can you think of any take-aways that you found particularly interesting?

While these questions seem quite prescriptive and straightforward, the participants often had plenty of feedback and points to make about each section. Additionally, if a participant made an interesting point, then I would ask specific follow up questions, which resulted in each interview being a rich and engaging discussion about the workshops.

3.4 Analysis

The recordings of the interviews were uploaded to Microsoft Stream, where they are stored, protected, as my university account had sole ownership and viewing permissions. I used an in-built MS Stream function that generates automatic captions of uploaded videos, to download a raw '.vtt' file that contained the captions, along with timestamps and other metadata. I then used a bespoke Python script, provided by COTADS team member, Sam James, that stripped the metadata and converted the file into a '.txt' file containing the captions.

Although the clinicians were members of the COTADS team, they have been included in the analysis, as they meet the criteria of non-experts with programming. They were not present for the co-design workshops conducted with young adults and family carers, and they were interviewed in the first round of exit interviews, at the same time as I interviewed the COTADS team.

I conducted a deductive thematic analysis [20] on the data corpus. I followed the six-phase process outlined in Clarke and Braun's paper:

1. Familiarise oneself with the data
2. Generate initial codes
3. Search for themes
4. Review themes
5. Define and name themes
6. Produce report

Armed with a (mostly) accurate automatic transcript, I then watched through the recordings, cross-referencing and correcting any errors in the script, for each interview, to produce an 'intelligent verbatim' [21] transcript. This process took a long time, but it was the first phase towards performing a thematic analysis – *familiarising myself with the data*. I also noted any interesting interaction data that could be observed from the recordings of screen shares from participants.

I used the qualitative data analysis software, NVivo, to thematically code the data corpus. My codes started as broad categories, but became narrower and deeper topics as I iteratively analysed the data. In writing the report, the themes that were established and named became the sections within my findings.

4. Findings

4.1 Opportunities of Jupyter Notebook for co-design

4.1.1 Notebook is a powerful tool for starting and directing discussion

The elements of the Notebook keep the conversation on track and provide plenty of stimulus to allow fruitful discussion that goes in many different directions. The visualisations provided by the Notebook are powerful tools as they provide a prompt for the participants to talk about, which was much more effective than discussing a topic without something tangible to compare it to.

> *"[without the notebook] it would have taken a lot longer, there wouldn't have been a focus on where we were going, and because we were such a large group from different professions, I think it would have been really hard to have our time to have those discussions but also summarize it. I think having that information there in front of us focused us and focused the discussion." (CLI3)*

The visualisations allowed participants to discuss the details of a specific example, which then raised questions and discussions about the applications and implications of the example within a general setting. For instance, in the first session, participants discussed a line graph that showed blood glucose data from a Continuous Glucose Monitor (CGM) machine for three different days. YA5 reported that "*having a prompt to look at and discuss is a lot better than just saying, 'what are your thoughts on a line graph?'*" By sharing a visual prompt, participants were able to refer to specific parts of the example and produce insights more easily than if asked to just discuss their thoughts about blood glucose data. The examples were used by participants as "*jumping off points*" (YA2) to ponder deeper questions and discuss what wasn't shown in the examples. Allowing participants to applaud and criticise the examples fostered broad, fruitful and considerate discussion.

> *"I think it provides a focus for the discussion. It raises questions, people think: is this doing the right job; is this a good way of visualizing it; are there other features that aren't being included? I think all of those are good questions about diabetes care and about machine learning." (YA2)*

Having fruitful discussion was integral to developing understanding. YA1 pointed out that group discussion gave them time to think deeper on a subject and further develop their own ideas as a result.

> *"I think it was nice to be in a group rather than individual because sometimes you could sit back and properly think about it. […] the discussion probably brought more out than if you had just asked me on my own, because sometimes what they were saying helped to push me to say more." (YA1)*

4.1.2 Notebook facilitates individual ideas

The notebook afforded participants to formulate individual ideas by working through the activity by themselves, before group discussion occurred. Although the members of the workshop were referencing the same content, everyone having their own notebook allowed space for original and uninfluenced thought, which could then be brought to the group for further evaluation. YA6 reported enjoying the fact that they had a "*private*" environment "*to test those ideas before sharing*", without a fear of "*judgment*" of the way they interacted with the activity. When faced with interactive activities, participants had the freedom to input their own ideas, such as personal parameters in the risk score slider, which enriched their personal understanding of an activity. Private use of the notebooks was complemented by the fact that participants had different lived experiences of diabetes, so their personal

parameters may have been different to other users', thus they drew conclusions from an activity as an individual, prior to discussion.

> *"I thought it was interesting, I put in my parameters on it, and saw what I would have." (YA1)*

By affording the user more agency in their comprehension of the handout, they had a deeper understanding of the concept at hand before discussing with the group. This was a powerful affordance because it meant that participants were not bound to the pacing of the group and could take their time with an activity if desired, which would not be the case if, for example, the group was following a slideshow presented by the host.

> *"I there's just something nice about giving each individual the capability of, in their own time, flicking through each day and having to think about it." (CLI1)*

Participants also identified that there would be less potential for them to impart their own ideas onto an activity, had they been sharing one notebook as a group, which would have limited their creativity and ability to explore the activity.

> *"I think there's a potential that people could be overly dominant, if it's just one notebook, […], I think that there is a possibility that people wouldn't be as creative in trying out ideas." (YA6)*

4.1.3 Notebook facilitates co-design

In addition to facilitating individual ideas, as stated in section 4.1.2, the notebook facilitated co-design through the sharing and discussion of the individual's ideas. Participants' personal understanding of diabetes and the applications of machine learning was enriched by the combination of notebook elements and group discussion.

> *"This was one of the points where the group nature of the workshop, combined with the notebook, helped me make realizations about my own condition and about how to design for a condition like diabetes." (YA2)*

Participants were overwhelmingly positive in describing how the group discussion about the notebook content furthered their own understanding of the task, and how sharing ideas helped to inform their own activities, such as YA3 describing how they were inspired by other participants' feature importance graphs, and how they used other people's ideas to improve their own graph's features.

> *"I thought it was really insightful to see other people's priorities, because sometimes people have very different priorities in terms of their condition […]. It helped me fill out my graph a bit more - there was things I didn't even think I wanted, and I was like, 'oh, that's interesting, I'll put that in mine.'" (YA3)*

Participants were reassured that they were on the right track with their activity when sharing ideas with the group. Since their child had only been diagnosed with diabetes for a year at the time of the workshops, FC3 reported concerns that they initially felt that they didn't have the knowledge to complete a feature importance graph. They felt that they were missing something that other family carers, "*who's kids had been diagnosed like five or seven years*," would have, and that the other participants were "*in a different place and they have a better understanding of [diabetes]*". Thus, when the participants were sharing their graphs, and FC3 found that "*despite so many different levels of experience, people were pulling out the same things*," they were reassured that their opinion, experience, and position as a participant in the study, was justified.

4.1.4 Notebook enables a step-by-step learning process that is key to reinforcing Notebook usage and understanding ML concepts

Due to the layout of the notebook and the session-based nature of the workshops, we created an environment that encouraged a step-by-step approach to learning ML concepts, and how to use the notebook itself. While they were separate learning outcomes, they do appear to be intertwined, as a

positive user experience with the notebook often resulted in a positive understanding of the ML concept being taught.

In terms of developing understanding of machine learning concepts, multiple participants identified that the notebook "*built up into thinking what would your ideal model be*" (FC2), and that this learning process meant that "*people really were engaged in thinking of everything they could that would possibly have an influence*" on their design for the final task, the feature importance graph. Since the final task produced the most important empirical data, it was vital that participants had a thorough understanding of ML before attempting the activity. YA4 described how the explanations provided by the notebook elements throughout the sessions had supported their understanding of ML applications to diabetes care. They reported that completing the previous tasks was important, in order to consolidate their understanding, before moving on to the final task.

> *"I think it built up quite nicely. [The machine learning model explanations] helped me to think about what sort of inputs would go into [the model], and how it was important to ensure that that was right before starting out with a big model that could be clinically used. […] I think I would have got used to being thrown straight into machine learning, but I guess it was useful to start more basic." (YA4)*

The other aspect of step-by-step learning was building confidence in the use of Jupyter Notebook itself. For those who had no prior knowledge of coding and JN usage, it was important to teach the participants how to interact with the notebook and internalise the process of running code cells first, to not overwhelm them when introducing the more confounding coding task. FC4 explained that their first impression of the notebook was total unfamiliarity, and that the notebook appeared "*alien*" to them. They reinforced the finding that introducing the coding task too early would have overwhelmed participants, as it would have been too many new things to do in one session, and they appreciated the "*gradual*" approach to using the notebook.

> *"[The notebook itself] was initially quite an alien screen to me - it's not the kind of thing that I would normally work with - so the way it worked was completely unfamiliar. Even just clicking on a cell and pressing play, I needed to learn how to do all of that, so if [the coding task] had been quite early in the process, […] typing in and adding things to it maybe would have been a couple of steps too far - it needed to be a bit more gradual." (FC4)*

After the first few encounters with code cells, participants became familiar with recognising and executing pre-existing code cells, which allowed them to focus on learning to code upon first contact with the coding activity.

4.1.5 Notebook facilitated a calm approach to controversial topics

The workshops approached controversial and potentially difficult to discuss topics surrounding diabetes. These topics included: socioeconomic and ethnic inequalities in diabetes care; the roles of metrics, such as BMI and HbA1c, in diabetes care; individuals' choice of technology, whether they use continuous glucose monitoring (CGM) technologies, such as 'DexCom' and 'Libre', as well as insulin pumps or manual injections; the individual nature of diabetes, how experiences and treatments of the condition differ from person to person, such as struggling with blood sugar control. Many of these topics can be embarrassing to discuss.

> *"I think [the dataset activity] was a really good way to sort of hammer down the point that everybody's diabetes and their diabetes control is so vastly different, with people using different tech as well. I think it was a really good conversation starter for that, because of myself not using any sort of technical sensors, closed loop systems, or a pump; and then other people do fingerprick tests and [have] a pump, or have a Libre and are still on multiple daily injections." (YA6)*

The notebook helped to facilitate the discussion of these sensitive topics by clearly displaying information, and then inviting open discussion, rather than railroading the conversation down a specific narrative path.

> *"I think it was really well communicated. It was really clear that making false assumptions and making decisions, particularly when it comes to medical issues and clinical need, based on a data set that doesn't represent a particular subset, and applying that decision to that subset is not just unhelpful, but potentially harmful." (FC4)*

Additionally, participants would find familiarity not only in the way that data was presented as charts and tables, but in the content of the data itself. For example, the blood glucose data graphs in the first activity were reminiscent of the graphs produced by CGM devices, which some participants were familiar with. As YA1 noted, "*It's quite similar to some of the ones that I'd get from the Libre, which is what I use now, which I find quite useful. I find it good that it's continuous*." YA6 appreciated how the graph displayed a 'bad day' (a day of high blood sugars meaning poor control), and that having a '*model*' graph that reflected a darker time of diabetes control meant that "*there isn't any embarrassment"* when discussing this topic, but in fact *"there is commonality*" between participants. It is an incredible positive that participants feel empowered by the content contained in the notebook and feel comfortable addressing sensitive topics.

> *"I think the way that individuals opened up individually meant that everybody felt pretty comfortable sharing their different perspectives. I think these three charts definitely facilitated and opened the conversation because of the way it represented things that people recognise from their lives." (FC2)*

4.1.6 Notebook supports and blends multiple learning styles

During the exit interviews, many participants reported that they enjoyed the fact that the notebook catered to multiple 'learning styles'. Despite this phrase not being part of the questioning, multiple participants used the term "*visual learner*"(YA3, FC3) to describe themselves or others, while other participants suggested ideas of "*implicit*"(YA5, YA6) and "*explicit*"(YA5, YA6) learning.

YA5 identified that the feature importance graph "*wasn't interactive and it was a bit more explicit rather than implied*", unlike the risk score slider which facilitated implicit learning due to its interactive nature. YA5, who had prior knowledge of ML as a computer science student, expressed that they preferred the explicit feature importance graph because "*it was a bit more in depth at showing that these models can be made of plenty of different features with various weightings."*

My findings indicate that it was important to have a blend of both implicit and explicit teachings, because the learning styles complement each other and allow for a broad, but also deep understanding. It was important to first establish a baseline understanding through implicit learning, before explicitly '*digging into it*'(YA6) by looking at more detailed, complicated examples that pushed the boundary of the participant's understanding. As someone without prior knowledge of ML, YA6 explains how the interactive, implicit activities provided necessary foundational knowledge, before that knowledge could be expanded on by studying more explicit information.

> *"I thought that it was important that we had the implicit one first, because that was the simpler one to understand. Had we jumped straight into the explicit, I think that it would have been, "how does this work?" and not understanding that there is more to it. Having the implicit first to understand how it works and then digging into it was the right approach for me." (YA6)*

The 'learning styles' that participants alluded to, are referencing Neil Fleming's VARK model [22], [23]– a popular, common knowledge model, that is often used to describe how people learn through different methods. VARK stands for 'Visual, Aural, Reading/Writing, Kinaesthetic'. The notebook supports this learning model by blending the different elements of learning: Visuals, through graphs and diagrams; Aural, from the workshop leader dictating and presenting the notebook, and from group discussion of the content; Reading, through the text contained within the notebook; and Kinaesthetic, from the 'learn by doing' interactive elements of the notebook. CLI3 praised the way that the notebook incorporates and caters towards these different ways of digesting information, "*I think different ways of putting information together suits different people differently*". YA3 identified themselves as a visual learner, and described how the visual information combined with the aural aspect of discussion, to give them the best possible understanding.

> *"[The notebook elements] obviously gave me visuals, which I said I learned by […]. I found the combination of using the notebook, and then speaking with the team and obviously the [HCI researchers] who were leading the sessions, the combination of the two helped me fully understand it and then I was able to give the best answers I could." (YA3)*

Even though FC3 identified themselves as a '*reading person*', they still stressed the importance of including visual representations to support understanding.

> *"without anything visual - and I'm not a visual learner, in all honesty, I'm a reading person - if you'd [explained] to me a machine learning model without those things, I would have been [sic] "What on earth are you talking about?" It would have made literally no sense." (FC3)*

Although no participants identified themselves as 'kinaesthetic learners', there is evidence that participants benefitted from the kinaesthetic, or 'tactile', learning elements present in the notebook. When describing the interactive risk score slider, FC2 enthused, "*I'm a sucker for things you can slide around and see the impact of, it's almost like gaming, isn't it, which I think is one of the engaging elements of using the notepad approach.*" In addition, YA4 verified that a *learning-by-experimenting* approach complemented their learning methods.

> *"I'm quite methodical in the way I think about things, so I quite enjoyed keeping different parameters the same and then changing other ones. So I quite enjoyed that from a 'how I work' point of view." (YA4)*

Interactive activities fulfil the criteria for hands-on, experimental learning, and participants were incredibly vocal about how the interactivity of the notebook was the standout feature; the aspect that elevated it when comparing JN to other presentation tools, such as a PowerPoint slideshow.

> *"The outputs of what you got, I thought were really excellent. The way that you could do things that are more than just showing a spreadsheet or showing a PowerPoint, the interactive models where we could enter parameters and see the changes ourselves – I thought that was great." (FC2)*

By combining the four aspects of the VARK model, Jupyter Notebook enabled a broad range of non-expert users to understand ML and thus participate in co-design centred on a difficult topic.

4.1.7 Notebook allows for the harnessing of coding by non-experts

Once the initial boundaries of approaching code were overcome, the harnessing of coding within JN was a powerful tool that enabled rapid development of ML ideas in a relatively simple manner. In the coding task, participants had to design their own feature importance graph, where they chose features and corresponding weightings. FC3 described how they struggled, at first, to assign numerical weightings to features because of their fictional nature. Upon initially coding values and generating a graph, FC3's understanding was rapidly developed by the visual representation of their figures; they could now visualise how the model's features were weighted, relative to each other, and iteratively adjust values until it looked right to them.

> *"When I initially did the features of importance, I remember thinking all these are just pie in the sky figures, not really sure, and I went back and found I could change them quite easily and quite quickly as well." (FC3)*

While they noted that the coding was not necessarily the most intuitive way of creating a graph, FC4 also described the benefit of the instant feedback from changing the code.

> *"It's the visualization of it that when you type it in it, it shows the impact of changing the value immediately. Whereas I think other ways are more intuitive in terms of yeah, we all know how to fill in a form or a Google doc or whatever, but then you wouldn't necessarily see what the graph then looks like until you have done it all, and then you'd have to go back and start all over again." (FC4)*

YA5 also identified the accessibility to Python and a coding environment that JN provides.

> *"I do like using a notebook to at least show code to people, because I always find a notebook is a lot easier than if you just had a Python file and having to get new people to set up everything for that." (YA5)*

4.1.8 Notebook supports accessing coding for the first time

Participants who were experienced at coding, such as YA2 and YA5, found the complexity of the coding to be trivial. YA5 reported that they "*found it very easy*" and didn't have "*any troubles with writing it or anything*," and YA2 stated that they had "*been coding for a while, so this was this was fine, it even did the plotting for me*."

However, presenting the Jupyter Notebook, a technical tool, to non-expert users was understandably difficult and met with apprehension. Participants, such as FC4, who had no prior experience with coding, described their first impressions of the notebook as "*daunting*", and FC1 expressed, "*when I opened it for the first time I was maybe worried a little bit and not sure how I was going to cope with it.*" However, with guidance from the workshop leaders, and by approaching the notebook slowly and with the step-by-step principles described in section 4.1.4, participants who were initially put off were still confident in giving the notebook their best attempt. Although FC1 was initially overwhelmed, when given the appropriate support to use the notebook, they reported, "*I found [the programming activity] easy after [HCI2] explained what to do, it was easy and user-friendly.*" YA6 also described the importance of having an instructor in the workshop to help when coding, "*it was really good that we had a supervisor in to nudge us in the right direction if we needed help with it.*" Likewise, FC4 reiterates that by the time the coding task came around, they found the task "*relatively straightforward*" and although they did make mistakes while creating code, they persevered and managed to succeed due to the gradual approach to the task.

> "*I think if this had come further up in the workshops and it was in the first or second [session], I think I'd have really struggled. It was better being in the last one and it was a little daunting, but relatively straightforward to do. I think I got tripped up and I don't know if I was the only one tripped up by adding in features and getting caught out by not putting commas or quotation marks in the right places.*" (FC4)

It is worth noting that in the workshop, all participants were able to successfully edit the example code and produce their own feature importance graph. The greatest positive finding was the reactions of participants who had never coded before, after succeeding in the task. Upon their first impression of the notebook, FC3 noted frustration and despair at the perceived technicality of the environment, "*when I first saw it, I thought 'Oh hell, I'm never going to get to grips with this,' because it looks very scientific to me and a bit geeky, computer science-y, and nerdy.*" However, when eventually tasked with writing code, FC3 discovered that the task was not as difficult as they first thought, "*I was just following the steps and when I tried it the first time I thought, 'oh I can do it, that's good!'*" FC3 went on to report that the coding task ended up being their favourite activity of the whole workshop, and that they gained satisfaction by being able to code successfully.

> "*I actually quite liked doing that coding thing when it came down to it. I thought I was going to hate it, but I quite liked the interactivity of it […] I had quite a little bit of a sense of satisfaction about being able to do it. I know it's ridiculous, but I was like, 'Oh yeah, this is easy actually.'*" (FC3)

YA6, who had no prior coding experience, also reported similar feelings of empowerment and jubilation when successfully coding. Their success with their coding enabled them to produce the most comprehensive feature importance graph out of all participants, with fourteen features in total. Because the example code only had placeholders for five features, this demonstrates that they had an adept understanding of altering code and adding elements to an array. Their comprehensive feature importance graph also demonstrates that the coding activity was beneficial to the production of empirical data through co-design.

> "*For me I had a couple of issues at the start, because I wanted to do more factors compared to what was initially put into the code, but I felt like I did alright with it, and as I got into it, I*

picked it up pretty quickly with how to stick the right things in. […] [I felt] really, really proud of myself because I'm not techy. I can work a computer - IT is absolutely fine - but coding is something that I never thought I'd be able to do. Because I don't understand it, […] it is quite intimidating. Especially with a big lack of women in STEM, sometimes it's hard to pique that interest of, 'actually, it's not as intimidating as you think it is, you can do it."" (YA6)

4.2 Challenges of using Jupyter Notebook for co-design

4.2.1 Technical nature of Jupyter Notebook makes participants feel like they are not equal co-design members

It was difficult for some participants, particularly the family carers, to interact with Jupyter Notebook because of its technical nature. FC2 assessed that the notebook was not inclusive to their peers, which had a detrimental effect on the collection of empirical data, because users had to struggle with operating the software as well as engaging with the content displayed.

> *"If you want to be inclusive, make it easy to consume and not have distractions to get to the empirical stuff. The best phone apps that we have are the simple ones, all the hard work is done behind the scenes, and the participants just have to focus on the question that's been asked, rather than how they interact with the notebook itself." (FC2)*

Difficulty with aspects of the software, such as navigating the document and executing code cells, made participants feel like they were not experts. FC3 explained how the presentation of the environment, such as the code in cells, made them "*feel like a complete fish out of water with this notebook because of this coding that I can see that I don't understand*". Whilst the code in the cells had been abstracted away as much as possible, it was still unfamiliar to the participant to the extent that it had a negative effect on their experience, and consequently, their understanding. FC3 continues, "*If [the code] was taken out I'd have probably quite a different feeling about it*," and went on to discuss how the design of the notebook didn't seem to have much consideration for non-expert users.

> *"It looked very scientific and that did put me off, if I'm honest. I just thought some of the things that were on there didn't really make sense to me, and it felt like there hadn't been any consideration for the way that some of the terminology might make the user feel." (FC3)*

FC3 further criticised the visual design of the notebook, explaining that the layout generated by the markdown file was not "*very visually appealing*" and that "*it kind of looked like a textbook*." FC3 explained how this 'textbook-like' was off putting to non-technical users: "*Textbooks scare me a bit because I'm not a maths or science person!*"

Because JN is a tool for data scientists, its usage was primarily meeting the needs and expectations of one stakeholder, the machine learning researchers, but it fell short of meeting the other participants' needs. By not using typical co-design tools, such as a shared document or whiteboard, it strengthened the data scientists' abilities to produce and distribute information. This came at the cost of the participants' abilities to digest and understand the information. FC4 discussed how JN was harder to use than other collaborative tools that they had used before, because of the technical nature.

> *"It wasn't particularly familiar. It was very different to anything I'd used previously. For someone who doesn't have any kind of coding or particular IT expertise, it did feel more technical than a lot of the other ways in which you can get people to share comments. I've been to other online workshops that are more sort of focus group-like where there are virtual post it boards and that kind of thing,, which is easier to get to grips with - the computational workbook felt far more technical." (FC4)*

4.2.2 Presentation in JN led to misunderstandings in sessions

The notebook sometimes failed to communicate the intentions of activities. This miscommunication resulted in breakdowns in different ways, such as participants not taking away the correct learning outcomes from activities, not understanding the information being presented by an activity, and not performing an activity correctly. Whilst most of the notebook was successful at communicating its intentions, these occasional breakdowns must be reported on.

The notebook failed to communicate the context of the risk score slider activity and explain that it was an example. During a conversation with YA6 about the risk score slider, they reported, "*I was so frustrated with [the risk score slider] because it's just not representative of diabetes management. You could be testing six times a day, but if you don't act upon it and don't inject, you're still at a high risk of going into hospital. There needed to be three different outcomes: hospitalization for hypoglycaemia, hyperglycaemia, and DKA.*" YA6 had missed, understandably so, a bullet point in a list of text above the risk score slider that said, '*The model shown is an example of what can be done and does not represent diabetes education or advice.*' By missing this one line in the notebook, YA6 became disillusioned with the workshop activity. For YA6, the context of this activity completely changed from foundational learning about the concept of ML, to frustration at the perceived simplicity of the model and lack of awareness from the designers.

> *"I might have not picked up that this wasn't a fully-fledged thing at first, so I think that's on me." (YA6)*

YA5 also expressed how they were unsure if the risk score slider was an example.

> *"I wasn't really sure if this was a genuine model […], so I was slightly concerned because I'm not sure how good this model really is. I feel like just three features and a value between 0 and 1 wouldn't actually produce any meaningful results." (YA5)*

The intended learning outcomes (ILOs) of activities were also not explicitly communicated to participants. When asked how well they thought the ILOs of the risk score slider had been communicated, FC4 responded, "*Maybe not that well, actually. I think it was obvious to us what we were investigating and what feedback we were giving, but in terms of learning outcomes for us, no I'm not sure that was communicated particularly well.*" I asked FC4 this question after YA6 reported their frustration at the risk score slider accuracy and depth, and their response reinforces that more care could have been taken to ensure that participants understood what each activity was trying to teach, in order to gain the best understanding possible.

> *"Now you've put it like that, that those elements were designed to increase the understanding, I get it. I don't think necessarily it was clear enough that that was the purpose of the notebook. I think it almost needed a little bit of pretext or something initially with a more basic explanation, and then an explanation of exactly what you've just said of, "we get that people taking part may not have any background or understanding of what machine learning is, it's obviously quite complex, but this is what it is in basic terms and the notebook is designed to replicate it to a point so that each of these elements show you how a machine learning model will work." I think maybe that context was missing at the beginning." (FC4)*

For YA1, the notebook failed to explain what the goals and intentions of the workshop were. Because of this, they felt that they could not make contributions to the best of their ability. YA1 needed the notebook to go further when explaining what contribution it expected from them, to feel encouraged to participate fully.

> *"I think it would have been helpful if someone had been like, "We are trying to achieve this" so I could have known what is helpful to you. Because if you had a full idea and you were really far in and then I was like, "this is rubbish" I would have felt really bad so sometimes I didn't know whether it would have been an appropriate time to say "I don't think this would be helpful" because it's your research. […] I would have been able to contribute more with a greater understanding of the project and where it was trying to go. I wouldn't say that I held loads back,*

but sometimes I felt like what we were discussing was a bit vague. I didn't really know what to say and what not to say." (YA1)

FC3 explained that the compressed layout of information in the notebook made it harder to digest information.

"Typically when you're writing a document for academic purposes, it would not be so squashed, like part 2 I would expect a new page. […] I'd definitely like new pages where a different part starts." (FC3)

4.2.3 Varied perceptions of the complexity of notebook content

When asking participants about the complexity of the ML content in the notebook, I received a variety of responses. The notebook provided simple explanations of ML concepts, and while some participants were satisfied with the amount of detail that the notebook went into, a handful of participants expressed interest in more complex explanations.

CLI3, an advocate for the given level of detail, reported, "*I think anything more perhaps would have been a little bit too much, I think it might have lost my focus from the information that I was there to provide.*" YA1 recognised that there was more to be said on the topic of machine learning, but deemed that the notebook had struck a healthy balance of detail and simplification.

"There's obviously a lot more that goes on, but I don't know what that is and how you'd visualize that, so for me personally, that's a decent level of detail and explanation, but also a good level of simplification because I feel like if you put too much maths in it, I wouldn't understand." (YA1)

While the development team wanted to keep the notebook elements simple to facilitate understanding in all users, FC2 thought that making the content of the risk score slider more complex, by adding more variables, would be insightful: "*With just the three dials, it is pretty easy to get to the bottom of what you're trying to say. Would I have got more or less insight out of having more or less of these [variables] to choose from? I'd have probably gone for more just to see.*"

YA5 observed that the simplicity of the examples could be counter-intuitive and create a narrative that machine learning was not capable of handling more than a few parameters, because none of the early examples incorporated more than 3 variable inputs. This would cause participants without prior understanding of ML to underestimate its power. YA5 suggested that some examples, particularly the one explained by the guest ML expert, could afford to be more complex.

"In the explanation of a model, this training and testing model is very simple as well. Because this isn't very interactive, […] it can be a bit more of a complicated image of a model to show the extent of them; rather than 'OK, these models are always just taking in one, two or three features, and any more than that, it might not be able to handle.' I feel like people who wouldn't be comfortable with machine learning wouldn't know [otherwise] if they haven't touched it at all before. I would assume it's quite obvious, but it was only because all the examples in the notebook are very simple, that I wasn't really sure how simple the actual models potentially were." (YA5)

FC2 also considered the effects of simplifying the explanations even further, by removing the teaching about ML completely. They questioned whether removing the ML context would still result in the same data empirically, because participants would still be asked the same questions, regardless.

"I do wonder if the focus on the machine learning model itself could have been a distraction. If you imagine that conversation, but without the machine learning bit thrown in, exactly the same questions that you did ask, you might have got more directly relevant insight as opposed to trying to bring people up to speed quickly on what a machine learning model is when you're still going to essentially ask the same questions at the end, which is 'what are the most important factors to you?'" (FC2)

While most of the participants that were eager for more detailed content had prior experience with ML, some participants without prior experience were still open to deeper ML explanations. When asked if they thought the explanations given in the workshop were sufficient, YA4 recognised the benefit of more information:

> *"I think in the workshop, perhaps because it's not like a fully-fledged finalized model that's going to be harder to do and I understand that, but I think maybe a little bit more information is a starting point. Just the basics of how these things are integrated would be useful, but I guess without a complete understanding of how a model is going to work, you can't really give a lot of information." (YA4)*

The 'one size fits all' notebook that was "*clearly designed everything for a non-technical audience*" (YA2) did not anticipate that some participants would have a deeper understanding of ML, and it did not cater to these participants.

4.2.4 Usability issues of the notebook

Participants reported that they would have found benefit in some training prior to the start of the workshops, to help them get used to the notebook.

> *"In terms of using the notebook, I think again, looking at age demographics, class demographics, people who wouldn't have as much exposure to technology because of class and socioeconomic reasons, they might not have the same exposure and therefore not be able to use the equipment as adequately, but then again, it doesn't mean there couldn't be a short training video on it." (YA3)*

> *"[Could there have been] some sort of pre-first workshop run through, or […] a training session on the workbooks, so before you even start thinking about diabetes or anything - a bit like 'Scratch' - build your own bar chart or something like that. I think that would have been good to give people the chance to do that beforehand, I know we were tight for time." (CLI1)*

Participants struggled with running code cells in the notebook. The notebook had an issue where code cells would not execute on first load, and the notebook kernel required restarting. I observed from participants' screen share during exit interviews, that this was an issue for multiple participants. Young adults who had no prior experience with JN (YA3, YA4, YA6) demonstrated fluency with the notebook by restarting the bugged kernel without problems. However, in the exit interviews, family carers (FC1, FC3, FC4) lacked the confidence and agency to fix the issue on their own and required guidance from me to restart the kernel. A further breakdown in internalisation of notebook usage was apparent when FC4 pressed the refresh button on their browser instead of the refresh button within the JN environment. During the exit interview, in response to the cell not working properly, FC2 remarked: "*I think it's a perfect example of a bit of extra complexity that isn't necessary to get some nice-looking charts up.*" FC3 raised their frustrations with needlessly having to run code cells:

> *"What I didn't understand is why it was there in the first place. […] what is the purpose of [running the cell] when [the graph] could just be there in the first place? […] or could you have a button next to it "click on this" so it's a bit more easy to [use]." (FC3)*

4.2.5 Comparison with other tools

Some participants did recognise the benefits of using JN for the workshops, such as CLI1: "*I do not think we would have got as far without the notebooks there to help us through, I don't think we would have got as far with just using 'Miro' whiteboards or post it notes or whatever.*" FC2 also appreciated the unique insights that the activities in the notebook provided, that other tools could not provide:

> *"The outputs of what you got I thought were really excellent, so that the way that you could do things that are more than just showing a spreadsheet or showing a PowerPoint, the interactive models where we could enter parameters and see the changes ourselves – I thought that was great." (FC2)*

However, many participants made comparisons between JN and other tools that were less favourable to JN. CLI3 describes how they found collaborating on shared slides was more familiar to them and easier to interact with than the notebook.

> *"When we moved on towards the end of the sessions and used slides to edit, and we were all working on the same [page], I personally found that easier. The kind of coherence, working together, but again, that's probably because I'm confident to do that, I've never seen a note[book] before, so I wasn't confident in doing it that way, but I think that's probably again quite personal." (CLI3)*

When working on activities that had no interaction, FC3 reported confusion at the use of JN over other tools that they were familiar with: "*I thought, 'I don't understand why he's using this notebook thing, why wouldn't you just use a PowerPoint or just share a document in that way?'*" YA3 reinforces FC3's thinking, when discussing the blood glucose graphs, which had no interaction besides changing the day, YA3 asked: "*these sort of graphs, why couldn't you just make them in Excel or something?*" When the use of JN was less justified, participants were more likely to complain about its use.

YA2 discussed that notebook offered a different type of collaboration to traditional tools such as a shared document, and that there was merit to both forms depending on the usage.

> *"What functionality do all of these methods of shared collaboration offer? I would say that it's different kinds of collaboration. It was the same notebook being used by everyone, but it wasn't one notebook being edited by several collaborators at the same time, it was just copies of the same notebook, whereas if you wanted to draw on ideas together, write things together, collaborate on a single graph together, you wouldn't use a notebook, you would use Google Docs [or] some other shared document platform like PowerPoint, but that wasn't really what people were [doing]. I guess that's why you guys switched to the PowerPoint exactly to do that, when you were asking people for their opinions and collating them. I think you chose the right solution for each part. […] this is more about interacting with code and the little plots, I think that the notebook is a simple way to do that." (YA2)*

4.3 Impacts on the use of Jupyter Notebook in an Online medium.

The notebook was used in an online medium, facilitated by video calling via Zoom, and there are interesting observations to be made about how the online nature affected notebook usage.

CLI2 considered the use of JN when collaborating remotely. They suggested that the notebook would not be as necessary in person, but still valuable. They discussed how the experience of using the notebook for co-design could potentially be augmented when in person, by blending notebook use with traditional physical co-design practices, such as sticky note sharing.

> *"I think when you collaborate remotely, the notebook is in invaluable. I think if you did it face to face then there's an option of not having a notebook, but I think it would be very hard to explain this without the notebook. I guess there are certain elements you could do without the notebook, I suspect. Or you could do notebook plus some interactive sticky note kind of stuff, but you still would need the notebook even if you did go face to face, I suspect." (CLI2)*

YA3 explained that they found it easier to present their model to the group than if it had been in person: "*I think I preferred it being online anyway because it's just too nerve wracking to do it standing up in front of peoples, it's easier to share your screen.*" However, YA5 found the presentation process disruptive because of the "*stop-start*" nature of participants sharing their screen: "*It's quite a tedious process to have people share their screens and then un-share and wait for someone else to share, rather than a collaborative whiteboard environment where it's a lot more flowing, rather than just blocks of people. It wasn't particularly bad; it was just a bit stop-start.*"

YA2 also explained how the added requirement of participants to share their screen was disruptive: "*What happened in the workshop, is that everyone in turn shared their screen and showed their features and their model because that's the only way you could do it - and that is kind of a limitation.*"

The online nature also caused the process of debugging coding errors to be more disruptive, as YA2 noted: "*'Share your screen. Let's go and look at that.' […] I think if you can kind of avoid [this process], that would just make life slightly easier.*" (YA2)

FC3 reported that they felt it was harder to quickly ask for help because of the online nature of the workshop: "*I don't even know whether I'm in the right place initially. I could have asked a question in real time, to say "Can you just check that I'm in the right place?" but I had to come back to zoom in order to do that.*"

Additionally, YA6 pointed out that the text chat feature of the online workshop was underutilised and could have had a positive impact on the discussion.

> *"I guess facilitating more use of the chat on zoom, the written chat. I don't think anybody used it other than to apologize for being late, whereas actually that would have been a really good tool to chip in or have thoughts, or people saying, 'what do you think about this?'" (YA6)*

Finally, FC4 explained that online meetings make it hard for everyone to have a say; more work could be done to make sure all participants are afforded the ability to contribute to the discussion.

> *"I was probably just as guilty of this as others, but I think there were some who talked more than others and I think online particularly, it's quite hard to make sure everyone has an equal voice." (FC4)*

5. Discussion

This project has investigated the feasibility of using Jupyter Notebook as a tool for co-design. The findings indicate that there are plenty of opportunities that JN provides, such as the facilitation of individual and collaborative ideas, and the blending of learning techniques to slowly build understanding of technical topics in non-experts. However, the findings have also identified challenges of using JN in a co-design setting, such as the technicality of JN, and usability issues, that made it difficult for participants to engage as well as they could. We discuss two criteria that require deeper investigation in order to establish JN as a worthy co-design tool: Design suggestions for Jupyter Notebook; Future work with Jupyter Notebook for co-design.

5.1 Design Suggestions for Jupyter Notebook

As per the findings in section 4.2.1, learning to use JN is currently part of collaborative learning process, because of its technical nature. This can be detrimental to learning because negative user experiences with the notebook often led to negative understanding. Therefore, design changes must be made to make JN more straightforward, which will reduce usability issues. These changes can be distilled into two themes: functional design suggestions, and aesthetic design suggestions.

5.1.1 Functional design suggestions

This section discusses features and alterations that could be made to JN to bring it more in line with other co-design tools.

Whiteboard / Post-it features

Participants made comparisons between JN and tools that provide a 'whiteboard' or 'post-it' note feature and made strong arguments for the benefits for collaboration that these other tools provide. If a collaborative idea sharing feature could be implemented into JN, it would combine the benefits of individual use that the notebook currently provides with the benefits of more interaction between participants.

Facilitating submission of notes and doodles to a post-it board would help users exchange nuggets of information and goes further to ensure everybody gets a say.

Allowing users to post their code cells to a post-it board could have also removed the need for screen sharing in the final activity when participants were presenting their own feature importance graphs.

Collaborative cell feature

Another possible feature would be using new technology for JN – real-time collaboration [24], [25]. A cell that updates in real-time for all users would allow multiple users to work on the same cell at the same time, and truly collaboratively design together. The current implementation of real-time collaboration in JN is experimental, and not quite suitable for our needs. It completely synchronises notebooks, rather than an individual cell, so the autonomy of individual notebooks would be lost.

Furthermore, this feature might be over the top, as a group working on an activity together could just nominate one person as a scribe, as long as everyone could see the cell in real-time.

5.1.2 Aesthetic design suggestions for notebooks

This section discusses aesthetic design alterations that could be made to notebooks to make then more accessible to participants using the notebook for co-design.

Improve presentation of information

The information notebook was observed to be too compressed, which made it harder to consume. The textbook-like nature of the markdown file could lead to information overload [26]. It is important that users do not miss critical information contained within the notebook because missing information in the notebook could lead to complete misunderstandings of the context of an activity. With some minor presentation changes to the notebook, information could be presented in a more digestible manner. Using text effects such as colour, font sizing, and emphasis, important information can be made to stand out more.

The notebook could have a cover page with introductory information. This page could contain reminders on the usage of JN, such as instructions on running cells, which could help participants who had forgotten, especially when returning to the notebook between workshops.

The notebook could even be made to be more like a book, with multiple pages to allow for easier navigation of the notebook. For instance, "turn to page 4" is much easier for participants to follow than "scroll down to section 4".

Each page in the notebook should also present key information about the session, clearly and at the beginning. Clearly presenting important information about an activity, such as the intended learning outcomes, context, and expectations, would help prevent breakdowns and misunderstandings, such as those reported in section 4.2.2.

Improve presentation of code cells

Code cells are a key component of the notebook, but they could be made more distinguishable to non-expert users, to make them easier to interact with. In JN, code cells are shaded light grey, but turn white when focused. If they were shaded a more distinctive colour and, it would be harder for users to miss them. It could also help with internalising the usage of notebook, as users associate the fact that they have to run the coloured cell. This intuition could be taken even further, for instance, by shading un-executed cells red and executed cells green, to give users visual feedback that they are doing the right thing.

5.2 Future Work with Jupyter Notebook for Co-design

5.2.1 Greater support for VARK learning styles can be facilitated.

Although the 'VARK' model's validity is shrouded in controversy [27], [28], it is still useful as a framework for discussing how different people learn and how the notebook complements these learning styles. In this discussion, support for the VARK model can be especially considered because participants directly alluded to learning styles unprompted, and how the notebook was beneficial in this aspect.

In this discussion we take a deeper look at how the use of JN supports learning through the VARK framework and discuss ways this could be pushed further.

The use of graphs and diagrams to present information is appealing to 'visual learners'. Further support for 'visual learners' could be provided by providing personal annotation tools that allows users to draw and doodle on their notebook.

For 'reading / writing learners' who digest information best through written text, it is important to ensure that the text they are presented with is properly framed, concise and informative, with important details highlighted if necessary. The 'Your Thoughts' notes section in the notebook further supports writers by providing them a space to jot down their thoughts.

'Aural learners' benefitted from the group discussion in the workshops, and from workshop leaders presenting and reading aloud elements from the notebook. The notebook could potentially do more do accommodate 'aural learners' – perhaps in the form of a dictation feature to read passages of text to the user.

For ‘kinaesthetic learners’ the interactive elements of the notebook are the standout feature. Whilst other tools for presenting information may be able to sufficiently cater to visual, aural, and reading/writing learners, they will struggle to match JN in terms of catering towards kinaesthetic learning. To further support kinaesthetic learners, maximising interactivity within the notebook is the best approach. However, the purpose of interactivity should be to enhance learning; care should be taken to not include unnecessary or misleading interaction.

Support for kinaesthetic learning through interaction is one of the best arguments in support of the use of JN in co-design. Facilitating kinaesthetic learning is the quintessential feature that other collaborative tools do not provide, and thus they are missing out on providing a vital way of digesting information and gaining deeper insight when tackling complex subjects, such as ML.

Furthermore, having a learning tool that is accessible for all types of learners can help with the diversity of the work group. Rather than a printout of text that only caters to typical academic (reading) types, varied representations of information opens co-design to a wide variety of people. For instance, if the usability issues are resolved, then there is potential to accessibly present information to groups such as people with learning difficulties, disabilities, hearing impairments, children, elderly people.

5.2.2 Aligning the direction of the notebook with the needs of non-experts

More work is needed in aligning the direction of the notebook with needs of non-data-science experts. Currently, the technical nature of Jupyter Notebook makes participants feel like they are not equal co-design members. Work needs to be done to empower people when using the notebook, so that the co-design principle of ‘everybody is an expert’ is reinforced.

By using a data science tool, participants must learn to share their expertise via the notebook as part of the learning process. Learning this new tool should not be a barrier, it should be a positive experience, so a further question that needs to be investigated when considering notebook as co-design tool is: in what ways can we make teaching notebook simple easy and enjoyable?

5.2.3 Not everything has to be as simple as possible

The notebook was designed for the lowest possible ability of participant. However, we found that this approach meant that participants with more ability lost faith in the workshop. With this in mind, there is an argument that the notebook content could have been made more technical, in order to engage the technical participants more.

Technical detail can be provided in the notebook, if it has sufficient explanation, and participants are aware that they don’t have to understand the more technical explanations. Arguably, providing extra information may be more beneficial than providing simple information that everyone understands. However, if this approach was implemented, care would have to be taken to consider how this makes participants feel. We must ensure that struggling participants are not disillusioned due to the complexity of the content. Potential remedies could include having expert ML researchers included in the sessions to provide insight into ML concepts and simplify ideas on-the-fly if necessary.

Another avenue might be setting participants homework related to the workshop, to elevate their understanding of ML. This homework could be optional but recommended if participants want to push themselves. On the contrary, if the homework was mandatory, it could present further issues and further disillusion participants.

When co-designing, the technical ability of participants is unclear until the workshop begins. Instead of producing a single notebook for all abilities, there is an avenue for designing bespoke notebooks for different levels of technical expertise. Participants may be more engaged by having a notebook that complements their ability. However, this would have implications for the distribution of knowledge and understanding as people would be operating at different levels, making it harder to run a group of

mixed ability in unison. Ultimately, this could lead to a reduction in diversity if participants have to be grouped by ability.

5.2.4 More coding

The overwhelming success of the coding task was unanticipated. While we expected most participants to be able successfully, the positive response and lack of problems with coding, bar a few teething problems, was completely unexpected. Furthermore, participants reporting that the coding task was their favourite, after first impressions of fear and hopelessness, is an indisputable success. As such, it is a shame that there was only one coding task, and we didn't push participants further. It would be very interesting to investigate how far participants with no prior experience of coding can be pushed in this environment. Maybe we just got lucky and had a great cohort of participants that were willing to go much further outside their comfort zone than could be expected.

5.2.5 Impact of virtual meetings affecting user experience of notebook

There is a space for discussion about the use of the notebook in an online vs offline setting.

While one might initially think that the notebook caters better to online workshops than an in person setting, this may not be the case. We observed that it was hard to debug and support individual users during activities online. It would be easier to help people in person, as a workshop leader could provide support to an individual without inadvertently addressing the whole group. In addition, it would be easier to observe and correct errors in code in person by observing the participants screen over their shoulder – there would not be a need to ask the participant to share their screen. Additionally, participants may feel more comfortable asking for help in person as they would draw less attention to themselves and feel less embarrassed.

Since the workshop has only used notebooks online, it would be interesting to investigate how else the use of notebooks would differ in person. In virtual meetings, it isn't possible to have an aside conversation with the person next to you – you have to talk to the whole group. Did this have a detrimental effect on productivity, and interaction with the notebook?

Other open questions left to future investigators include: Would the reception of notebooks be different in person? Should the form of the notebook change when presented in person? Would it be better to return to classic co-design tools such as printouts, slides and sticky notes in a physical setting? How would people physically sharing the same space augment notebook usage?

6. Limitations

A potential limitation of the study could be that it was sometimes hard to gain critical feedback from the participants. Because the participants were non-experts at using Jupyter Notebook, they may have felt unqualified or unjustified when giving feedback, especially negative opinions about the software.

Additionally, because of the amicable and professional relationship I had formed with participants during the workshops, and the personal stake they held in the workshop's success due to the diabetes context, they may have felt obliged to be overly positive or less pessimistic in their feedback.

Whilst I tried to address these potential limitations in my exit interviews, by reinforcing that all opinions are welcome, and spending plenty of time discussing the workshops, it may have been prudent to have an interviewer that was independent from the workshops.

Further to this, perhaps a deeper review of using Jupyter Notebook as a collaborative would be achieved if it was the primary objective of a workshop, since exit interviews were rife with empirical feedback and context-specific feedback that was not relevant to my study. The use of notebook was not the first thing on participants minds during the workshops and the exit interview, and therefore may have prevented them from being as introspective as possible on this topic.

Finally, the research group was not very ethnically, socio-economically, or educationally diverse. The group mostly consisted of white people educated to a university level, which could have also limited the findings of the study.

7. Conclusion

Jupyter Notebook was not designed to be a collaborative tool, but it can be used as a collaborative tool for co-design. The benefits to learning that JN provides when used in this way, are significant.

JN empowers the user to have individual agency and control over their own workspace in a collaborative environment, which allows the user to establish individual ideas and increase their understanding of difficult concepts. JN also facilitates interactive learning, which is a powerful learning device, and this distinguishes it from traditional collaborative tools that do not have these capabilities.

The challenges of using JN are not insignificant, however, and the quality of some collaborative activities, such as sharing ideas, suffers as a result. Changes and additions need to be made to the JN system to bring it in line with other collaborative tools.

As per my research objectives:

- I have identified opportunities and challenges of using Jupyter Notebook as a collaborative tool for co-design.
- I have discussed changes to the Jupyter Notebook used in the COTADS project, and future work for the general use of Jupyter Notebook, but further investigation into future work for Jupyter Notebook is still necessary – there is plenty more to discuss on the topic and work to be done to make Jupyter Notebook a more viable option of collaborative tool for co-design.

References

[1] C. E. a. A. T. Muhammad Aurangzeb Ahmad, “Interpretable Machine Learning in Healthcare,” *In Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB '18). Association for Computing Machinery, New York, NY, USA, 559–560,* 2018.

[2] T. S. N. A. Parikh RB, “Addressing Bias in Artificial Intelligence in Health Care,” . *JAMA. 2019;322(24):2377–2378.*.

[3] Z. W. C. B. a. S. O. April Yi Wang, “Callisto: Capturing the "Why" by Connecting Conversations with Computational Narratives.,” *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, 1–13.*, 2020.

[4] K. Nygaard, *The iron and metal project: trade union participation. Computers Dividing Man and Work: Recent Scandinavian Research on Planning and Computers from a Trade Union Perspective, 98.,* 1979 .

[5] Y. Sundblad, “UTOPIA - Participatory Design from Scandinavia to the World,” October 2010. [Online]. Available: doi.org/10.1007/978-3-642-23315-9_20.

[6] T. a. A. K. Zamenopoulos, *Co-design As Collaborative Research. Connected Communities Foundation Series,* Bristol, 2018.

[7] *Ioannidis JPA (2016) Why Most Clinical Research Is Not Useful. PLoS Med 13(6): e1002049. https://doi.org/10.1371/journal.pmed.1002049.*

[8] *Domecq, Juan Pablo et al. “Patient engagement in research: a systematic review.” BMC health services research vol. 14 89. 26 Feb. 2014, doi:10.1186/1472-6963-14-89.*

[9] Slattery, P., Saeri, A.K. & Bragge, P. Research co-design in health: a rapid overview of reviews. Health Res Policy Sys 18, 17 (2020). https://doi.org/10.1186/s12961-020-0528-9.

[10] Di Lorito C, Bosco A, Birt L, Hassiotis A. Co-research with adults with intellectual disability: A systematic review. J Appl Res Intellect Disabil. 2018 Sep;31(5):669-686. doi: 10.1111/jar.12435. Epub 2017 Dec 12. PMID: 29231285..

[11] J. a. J. A. S. Donia, “ Co-Design and Ethical Artificial Intelligence for Health: An Agenda for Critical Research and Practice. Big Data & Society, July 2021, doi:10.1177/20539517211065248.”.

[12] A. Ayobi, K. Stawarz, D. Katz, P. Marshall, T. Yamagata, R. Santos-Rodríguez and P. a. O. Flach, “(2021). Machine Learning Explanations as Boundary Objects: How AI Researchers Explain and Non-Experts Perceive Machine Learning. In: 2021 Joint ACM Conference on Intelligent User Interfaces Workshops, ACMIUI-WS 2021, 13-17 Apr 2021, College Station.”.

[13] K. S. D. K. P. M. T. Y. R. S.-R. P. F. a. A. A. O. 2. Amid Ayobi, “Co-Designing Personal Health? Multidisciplinary Benefits and Challenges in Informing Diabetes Self-Care Technologies. Proc. ACM Hum.-Comput. Interact. 5, CSCW2, Article 457 (October 2021), 26 pages. https://doi.org/10.1145/3479601”.

[14] “Perkel, Jeffrey M. "Why Jupyter is data scientists' computational notebook of choice." Nature 563.7732 (2018): 145-147.,” [Online]. Available: https://doi.org/10.1038/d41586-018-07196-1.

[15] JupyterLab, “Jupyter.com,” [Online]. Available: https://jupyter.org/.

[16] W. Stein, “CoCalc Whiteboard,” [Online]. Available: https://cocalc.com/features/whiteboard.

[17] “JDRF Information about Type-1 Diabetes,” JDRF, [Online]. Available: https://jdrf.org.uk/information-support/about-type-1-diabetes/facts-and-figures/.

[18] G. L. H.-C. L. R.-H. a. F. J. (. Sadler, “Research Article: Recruitment of hard-to-reach population subgroups via adaptations of the snowball sampling strategy. Nursing & Health Sciences, 12: 369-374.”.

[19] C. Duckworth and mjbonifa, “First release of COTADS Co-design notebook,” [Online]. Available: https://doi.org/10.5281/zenodo.6376782.

[20] “ Virginia Braun & Victoria Clarke (2006) Using thematic analysis in psychology, Qualitative Research in Psychology, 3:2, 77-101, DOI: 10.1191/1478088706qp063oa”.

[21] “verbatim vs intelligent vs edited transcription,” SummaLinguae, [Online]. Available: https://summalinguae.com/data/verbatim-vs-intelligent-vs-edited-transcription/.

[22] “Fleming, N.D; (1995), I'm different; not dumb. Modes of presentation (VARK) in the tertiary”.

[23] “Fleming, Neil D. Teaching and learning styles: VARK strategies. Neil Fleming, 2001.”.

[24] K. Jahns, “Jupyter Blog, How we made Jupyter Notebooks collaborative with yjs,” JupyterLab, [Online]. Available: https://blog.jupyter.org/how-we-made-jupyter-notebooks-collaborative-with-yjs-b8dff6a9d8af.

[25] JupyterLab, “Real Time Collaboration - JupyterLab documentation,” [Online]. Available: https://jupyterlab.readthedocs.io/en/stable/user/rtc.html.

[26] R. J. a. H. d. Poot, “Information overload: why some people seem to suffer more than others.,” *In Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles (NordiCHI '06).*, p. 397–400, 2006.

[27] “Husmann, P.R. and O'Loughlin, V.D. (2019), Another Nail in the Coffin for Learning Styles? Disparities among Undergraduate Anatomy Students’ Study Strategies, Class Performance, and Reported VARK Learning Styles. American Association of Anatomists, 12: 6-”.

[28] W. L. e. a. Leite, ““Attempted Validation of the Scores of the VARK: Learning Styles Inventory With Multitrait–Multimethod Confirmatory Factor Analysis Models.,” *Educational and Psychological Measurement, vol. 70, no. 2, Aation of Anatomists, 12: 6-*, 2009.