Marine Biology Research at Bahamas

Unique and exclusive partnership with the Gerace Research Center (GRC) in San Salvador, Bahamas to offer marine biology research opportunities for high school teachers and students.

- Terra has exclusive rights to offer the program to high school teachers and students around world.
- All trips entail extensive snorkeling in Bahamian reefs as well as other scientific and cultural activities.
- Terra will schedule the program with GRC and book the flights from US to the GRC site.
- Fees include travel within the US to Island, lodging, meals, and hotels for transfers, and courses.
- For more information, please visit terraed.org/bahamas.html

Terra is a N.Y. based 501.c.3 non-profit organization dedicated for improving K-16 education
Table of Contents
June 2022 | Volume 4 | Issue 3

01 Effectiveness of Tele-therapy from the Perspectives of Training Therapists and High School Students
   Brianna Son

06 Hierarchical Optimal Path Planning (HOPP) for Robotic Apple Harvesting
   David W Liu

13 The Rationality Analysis and Prediction of THE World University Rankings
   Yingao Gao

17 Simulation of Heat Diffusion in a Tidally Locked Exoplanet
   Giulia Murgia

22 Cancer mRNA Vaccines as a Promising Approach for Treating Luminal A Breast Cancer
   Alexandra Giuliani

28 Analyzing Consumer Behaviors and Attitudes Towards Plastic Bag Consumption in Hanoi Wet Markets
   Jinmoo Yoo

34 Identification of fimS Regulator Motifs in K. pneumoniae and E. coli to Advance Antivirulence Treatment of Type 1 Pili
   Sarah Joseph

39 Replication Stress: Cancer’s Double-Edged Sword
   Keshav Saigal

47 Differentially Expressed Genes and Pathways Involved in the Development of PTSD in World Trade Center Responders
   Kevin Toledo, Harmanpreet Kaur, Miguel A. Javiel, Jennifer Barreto
<table>
<thead>
<tr>
<th>Page</th>
<th>Title</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>55</td>
<td>Analysis and Comparison of Safe Organic Ant Repellents</td>
<td>Jinghan Li</td>
</tr>
<tr>
<td></td>
<td>and Travel Restriction Measures using Reinforcement Learning</td>
<td></td>
</tr>
<tr>
<td>69</td>
<td>Invisible Labour: Comparison of Unpaid Work Done by Women and Men</td>
<td>Neasha Mittal</td>
</tr>
<tr>
<td></td>
<td>across Developed and Developing Countries</td>
<td></td>
</tr>
<tr>
<td>76</td>
<td>Applications of Australian Native Aquatic Plants on Purifying</td>
<td>Nguyen Nathan</td>
</tr>
<tr>
<td></td>
<td>Wastewater Sources</td>
<td></td>
</tr>
<tr>
<td>83</td>
<td>Identifying Factors Affecting the Spread of Variants of COVID-19 by</td>
<td>Peter Ma</td>
</tr>
<tr>
<td></td>
<td>Regression Models to Inform Policymaking</td>
<td></td>
</tr>
<tr>
<td>96</td>
<td>Observing Brainwaves when Playing Games Using a Ruby-based Wireless</td>
<td>Ri He</td>
</tr>
<tr>
<td></td>
<td>Brainwave Measurement System</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>Gap Between Scientists and Lay Public Remains: Genetically Modified</td>
<td>Ryan Hadi</td>
</tr>
<tr>
<td></td>
<td>Food Knowledge Negatively Influences Opinions</td>
<td></td>
</tr>
<tr>
<td>105</td>
<td>Controlling the COVID-19 Pandemic with Masks and Social Distancing</td>
<td>Shuying Liu</td>
</tr>
<tr>
<td>109</td>
<td>DNA Origami: Its Evolution and Application in Nanotechnology</td>
<td>Kavi A. Ullal</td>
</tr>
<tr>
<td>116</td>
<td>Familiarity-Dependent Relaxation Responses of Classical Musicians</td>
<td>Veronica Li</td>
</tr>
<tr>
<td></td>
<td>to Music Selections</td>
<td></td>
</tr>
</tbody>
</table>
Effectiveness of Tele-therapy from the Perspectives of Training Therapists and High School Students

Brianna Son
Flint Hill School, 3320 Jermantown Rd, Oakton, VA, 22124, USA; briannason89@gmail.com

ABSTRACT: The prevalence of depression and anxiety symptoms have doubled among adolescents during COVID-19. The pandemic drastically changed the landscape of mental health services as stricter measures such as quarantines and lockdowns were put into place. Tele-therapy - virtual therapy sessions - became a more prominent source of treatment delivery. This study seeks to further examine the attitudes about the effectiveness of teletherapy services based on the experiences of training therapists and adolescent patients. Five high school students receiving tele-psychoeducational services and five training therapists from a doctoral program in Clinical Psychology were recruited. These participants each completed the questionnaire regarding the effectiveness of teletherapy. Overall, the flexibility of teletherapy was a major benefit mentioned by both high school students and training therapists. However, in-person therapy or the hybrid model was preferred by both the student and therapist participants as this form of therapy provided a greater sense of connection and alliance between patients and therapists. Future studies should expand on this initial exploratory work, with a greater number of participants, including patients with differing levels of clinical risk and therapists with additional experience.

KEYWORDS: Behavioral and Social Sciences; Clinical and Developmental Psychology; teletherapy; mental health; Covid-19 pandemic.
level of treatment satisfaction. In fact, studies displayed that the majority of patients receiving remote psychological services are very satisfied with receiving tele-therapy; teletherapy reduces wait times and offers an alternative and safer environment for those with anxiety.

Not only is teletherapy beneficial for patients, but it is also beneficial for therapists as well. It offers a great work-life balance, a greater sense of safety, and opens up new opportunities to work with patients from all over the world. Austrian therapists, for instance, were prompted to complete an online survey in order to assess their attitude changes towards teletherapy. The therapists indicated a high level of telehealth positivity, with an improved result during the course of the pandemic lockdown. However, this study also denoted the need for reimbursement policies and secure software solutions to continue and better this technique of online therapy.

Experts say that if you're in need of mental health services, getting them remotely — via technologies like audio or video call — is unquestionably better than getting no therapy at all. But is teletherapy as good as seeing a therapist in person? The limitations of teletherapy cannot go unnoticed. Teletherapy is not feasible for all patients or clinics as technology is difficult for certain people to use and requires high-speed internet, computer access, and privacy in the home which not all patients have available to them. Conducting a session through technology also limits the personal connection that therapists and patients can forge in person.

This information points to my research question of how effective teletherapy is on high school students diagnosed with mental health disorders and for trained therapists conducting these services.

**Methods**

The present qualitative study seeks to further examine the attitudes about the effectiveness of teletherapy services based on the experiences of training therapists and adolescent patients. Two questionnaires were developed: one for high school students receiving teletherapy services, and a second for training therapists conducting teletherapy. The questionnaires were then distributed via email to participants. Responses were analyzed quantitatively and qualitatively to assess how both patients and providers view teletherapy as a treatment option for a variety of mental health disorders.

**Participants:**

High school participants receiving tele-psychological services (N = 5) were recruited from a private high school in Virginia. The first two respondents were recruited intentionally, and the survey was sent out to three players on the high school soccer team. Training therapists (N = 5) were recruited from a doctoral program in Clinical Psychology in New York City. The mean age of high school participants was 16.4, with an age range of 15 to 17 and a grade-level range of 10th to 12th; 4 were female-identified, 1 was nonbinary. The mean training level range of the training therapists was 2 to 3 years; 1 participant was male-identified, 4 were female-identified. The survey was administered online from September to October of 2021. It is important to note that the survey was initially sent to 10 students, but only five responses were received, for both high school students and therapists.

**Materials:**

**Questionnaire for high school students receiving teletherapy.** The questionnaire for high school students receiving telepsychology services was designed with the goal of gathering attitudes about the effectiveness of teletherapy services based on experiences receiving in-person therapy and teletherapy. The online questionnaire was developed based on previous research on patient attitudes towards therapy. The questionnaire included 11 open-ended, yes/no, and Likert scale questions (i.e., How effective do you find teletherapy for you? (Rate on a scale from 1 (not effective) to 5 (highly effective)); What do you see as the benefits and limitations of the treatment you are receiving online? If given the option, would you prefer to receive therapy in person, online, or via a hybrid model? Why?; What do you see as the benefits and limitations of online treatment?).

**Data Analysis:**

Means and standard deviations were calculated for numerical responses. Themes were analyzed for short answer responses across participants. The short answer questions were reviewed initially participant by participant. Each of the questionnaires for the therapists and high school students was compiled into spreadsheets, displaying all of the responses in one document. One coder identified the common theme in each of the responses for every question then discussed it with the second.

**Qualitative Analysis for Therapists:**

The five therapists’ responses to the questionnaire were read then analyzed for common themes. Common themes were identified by one coder and discussed with a second; themes were based on each question of interest from the survey, and quotes were extracted from the transcripts. Themes were then classified into five categories: 1) limitations of teletherapy; 2) treatment with children versus adults; 3) teletherapy effectiveness; 4) attitude change toward teletherapy; and 5) in person, online, or hybrid treatment.

**Qualitative analysis for high school students:**

The five students’ responses to the questionnaire were read then analyzed for common themes. Common themes were identified by one coder and discussed with a second; themes were based on each question of interest from the survey, and quotes were extracted from the transcripts. Themes were then classified into three categories: 1) limitations of teletherapy;
2) benefits of teletherapy; 3) in person, online, or hybrid treatment.

Results and Discussion

Quantitative Analysis for Therapists:

On average, training therapists who participated in the survey reported seeing 6 patients (SD = 4.18). These therapists saw patients with an age range from 7 to 39 years old.

The therapists were on average 2.5 years into their doctorate degree and were participating in ongoing training courses (SD = 0.5). It was reported that on average, the therapists have been conducting teletherapy for 18.6 months (SD = 8.04). The longest treatment the therapists conducted online was 11.6 months, while the shortest treatment the therapists conducted online was 6 months. The therapists reported treating patients with a range of diagnoses - Depression, Anxiety, ADHD, Narcissistic Personality Disorder, PTSD, OCD, Recovering Substance Use Disorder, and Bipolar Disorder. The risk levels of the therapists' patients were low to medium (See Table 1).

Table 1: Demographic and descriptive statistic for training therapists (N=5).

<table>
<thead>
<tr>
<th>Description</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average number of patients</td>
<td>6</td>
<td>4.18</td>
</tr>
<tr>
<td>Average number of months</td>
<td>18.6</td>
<td>8.04</td>
</tr>
<tr>
<td>conducting teletherapy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average number of months longest</td>
<td>11.6</td>
<td>3.36</td>
</tr>
<tr>
<td>therapy case</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average number of months shortest</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>therapy case</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Limitations of Tele-therapy:

Limitations of teletherapy are noted by trained therapists such as loss of emotional connection and loss of intimacy with the patient. Overall, teletherapy brings challenges to the establishment of the therapeutic working alliance (the relationship between patient and therapist). As one training therapist describes:

There is a loss of intimacy and intensity that can be used in the treatment when it is in person. Some patients can seem less invested in the process.

Another therapist notes how treatment online can inhibit the patient and therapist's ability to explore core issues and dynamics. As the therapist reports, therapy via videoconferencing can allow the patient to hide and avoid challenging topics:

Additionally, I think it can be hard for patients to remain in a vulnerable/emotional space during the session like they can turn it off or move on from hard feelings easier because of the screen.

Practical obstacles are also noted among therapists as contributing to difficulties when conducting sessions virtually. Obstacles such as a lack of privacy, challenges with WIFI during sessions, the ability for the patient to turn the camera or sound on and off, and not occupying the same physical space damaged the effectiveness of conducting teletherapy. Nonverbal cues are absent during teletherapy, which causes disruptions in the process. Not being able to be in the same room as the therapists can make it difficult for the patient to be comfortable to fully disclose.

You sometimes don't even see what patients look like fully so hard to tell if they've lost weight or look more disheveled. Also, WIFI issues and an internet connection can inhibit treatment significantly.

Children vs. Adults Treatment:

Generally, treating children through teletherapy is harder than adults due to developmental impediments: children's attention span are shorter, which makes it easier for them to become distracted by their surrounding environment.

Moreover, since children's treatments consist in a large part of play, finding ways to adapt the play space to the online space can prove difficult, especially for less verbal children. Adults on the other hand, primarily engage with therapists through conversation, and they are more used to being sedentary for a longer period of time. Therefore, the nature of the interaction during teletherapy is still similar to their usual lifestyle, unlike that for children.

When children are less verbal, it can be more difficult to engage with them online. In a physical space, we can engage by using shared toys and the child can feel grounded and more willing to share in the comfort of a real-life empathic therapist.

Tele-therapy Effectiveness:

There is a 100% agreement among training therapists questioned that teletherapy is an effective method of service. Tele-therapy provides the patients with the necessary support and mimics the therapy experience well, especially for adults and older teens.

Even though it may be virtual, it is still a vitally important experience and relationship to have on a regular basis in order to process and make meaning out of our lived experiences.

Furthermore, there are a number of unique benefits of teletherapy that in-person therapy cannot provide. Virtual therapy sessions might make it easier for patients to show up on time, may foster more disclosure for those who become anxious about leaving their home, and make those more comfortable with this modality.

You can still build rapport and relationship, and it can even benefit patients who value their privacy or have certain pathology that lends itself towards being mistrust of others.

Attitude change toward teletherapy:

Overall, teletherapy is an effective method of treatment according to the training therapists questioned. Training therapists report being skeptical at first due to the concerns about whether or not connection will be present between therapists and patients; however, this method proves to be a compelling alternative. One therapist discusses her concerns when she began treatment and the realities of her experience:

I was really worried about whether I would be able to feel connected to my patients or be able to feel their emotions/mental state but I think that has proven easier than I anticipated actually.

Although teletherapy is a practical method, some therapists still preferred in-person treatments. There seems to be fatigue evolving from the Zoom setup. Since the pandemic has increased the use of teletherapy, the long hours of holding Zoom sessions could tire the therapists and patients as well. As one therapist notes:

I believe this fatigue stems from staring at a screen for too long and not having all the stimulation of the non-virtual world that would keep me energized and engaged.
For some training therapists, who are just beginning to see patients, teletherapy is the only version of the treatment that they are familiar with. Therefore, they do not have an attitude change because they don’t have an alternative that they can compare it to.

**In-person, online, or hybrid treatment:**

100% of the therapists prefer hybrid treatment – a mix of in-person and teletherapy. Tele-therapy offers great flexibility, in terms of scheduling and commuting for both patients and therapists. A hybrid model can prevent patients and therapists from feeling burdened due to its option of teletherapy. As one training therapist observes:

> I think some of my patients really wouldn't be able to make it to the session given the distance of the clinic to her home and if it makes it so that patients are more consistently engaged with the option of teletherapy I think that's great!

Therapists emphasize that in-person therapy is still more effective than being online. Especially for children, in-person therapy is a great option to have as it helps them engage better. A therapist remarks that:

> I would prefer a mostly in-person model with a small percentage of sessions being online. Overall, I believe that in-person therapy is more effective than online.

**Quantitative Analysis for high school students:**

The average age of the high school students was 16.4 years old (SD = 0.89). On average, high school students who participated in the survey reported taking part in therapy for 4.7 years (SD = 2.05). Half of the sample reported starting and stopping treatment within that time period. The mean length of time the students have been receiving therapy on an online platform was 20.6 months (SD = 1.95). Some of the students do group therapy online, while others do individual sessions. Students have received therapy in person for an average of 37.6 months (SD = 38.01). Students reported on a LIKERT scale (1 = not at all effective; 5 = highly effective) how effective they found online therapy (M = 3.6; SD = 0.55), and in-person therapy (M = 4.2; SD = 1.30). Based on the mean responses, students found in-person therapy to be more effective (see Table 2).

Table 2: Demographic and descriptive statistic for high school students in therapy (N=5).

<table>
<thead>
<tr>
<th>Description</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>16.4</td>
<td>0.89</td>
</tr>
<tr>
<td>Months in treatment total</td>
<td>4.7</td>
<td>2.05</td>
</tr>
<tr>
<td>Months in in-person therapy</td>
<td>37.6</td>
<td>38.01</td>
</tr>
<tr>
<td>Effectiveness of teletherapy</td>
<td>3.6</td>
<td>0.55</td>
</tr>
<tr>
<td>Effectiveness of in person teletherapy</td>
<td>4.2</td>
<td>1.3</td>
</tr>
</tbody>
</table>

**Limitations of Tele-therapy:**

Not being able to physically interact with the therapists is a common limitation for the students. Furthermore, receiving treatment in another environment besides their home can provide the students more freedom, but teletherapy eliminates this option:

> Along with this, I feel like getting out of the house/into a new environment could be beneficial for some.

Students also mentioned that distractions are recurring negative habits during the online sessions. Not being there physically leads to inefficiency in visualizing things for students and increases external distractions.

**Benefits of Tele-therapy:**

Flexibility is another common benefit of teletherapy that the students mentioned. Along with school and sports practices, students are offered great flexibility as they can receive therapy from home if they live at a far distance from their therapist. As one student comments:

> I don't have to leave as early from school practice if I need to get home for therapy, since I live 5 minutes from school but 45 from therapy.

Students are fond of the comfortable environment they are already familiar with when receiving teletherapy at their homes. Some of these students consider their homes their “safe space” and enjoy receiving sessions there.

> Some benefits of online therapy is I am already in my safe space at home/am comfortable.

**In-person, online, or hybrid treatment:**

60% of the students prefer the hybrid option, and 40% of the students prefer in-person therapy. Students would like the option of doing one session a week online and the others in person.

> I would prefer a hybrid model if I was doing multiple sessions a week (like 2), but I feel like it should be an option every week.

Face-to-face therapy is important for students and therapists to connect emotionally with one another. However, some students still prefer the hybrid model as it provides them with the flexibility to fit their schedules. Since distractions are mentioned as common limitations of teletherapy, students said that being in person reduces the distractions and increases the commitment to the treatment.

**Discussion**

As the results reveal, both the high school students and the training therapists highlighted the flexibility of teletherapy as one of the benefits of online sessions. Having the teletherapy option allows them to move around their schedule and create changes without affecting their individual plans excessively. The high school students and the training therapists’ responses accentuated the importance of in-person therapy. Neither the high school students nor the training therapists strictly preferred online treatment only; instead, they all either preferred the hybrid model or in-person therapy. A limitation both groups pointed out was the lack of intimacy within virtual sessions. Being online impedes the emotional connection that can be formed between the therapists and patients as they won’t be physically present with one another.

What can the results of this study imply for treatment going forward? This study shows the significance of teletherapy, but also its limitations of it. All of the participants in this study either preferred the hybrid model or in-person therapy. As the pandemic is slowing down, in-person therapy should shift to becoming more accessible once again.

Despite the significant findings, there were several limitations to note in the study. The first limitation was the small sample size. There were five high school students from
the same school and five training therapists from the same graduate school. This provides a very small sample of data and a lack of variety in responses. A greater number of participants would have provided more accurate results, in that this would lessened potential biases.

It is also worth noting that the survey was initially sent to 10 students, but only five responses were received. Perhaps the people that chose to respond felt stronger about the topic than the people who did not respond, leading to a non-response bias.

A second limitation was that the training therapists did not have experience conducting in-person therapy; this gives biased opinions of teletherapy as online therapy was the only form of therapy the therapists had experience with.

Training therapists included in this study focused on psychodynamic treatment - this could introduce bias. Psychodynamic therapy helps patients discover patterns, emotions, and thoughts to further explore themselves. Through a virtual session, however, it might be more difficult for this strategy to be conducted as several therapists and patients mentioned that there is a loss of treatment alliance as reported by patients and therapists through the screen.

This study could be performed with a greater number of participants, for both high school students and training therapists. This will provide a broader response and can lead to another questionnaire being drafted in order to get further information on the effectiveness of teletherapy. Another potential future study could use standardized questionnaires (Brief Symptom Inventory, or Outcome Questionnaire) or structured interviews with measurable variables to assess whether an adolescent’s mental health has actually improved over the course of teletherapy treatment compared to treatment in person.

■ Conclusion

The purpose of this research was to examine the effectiveness of teletherapy based on the experiences of training therapists and high school students. From the analysis conducted, teletherapy is a positive form of treatment due to its high level of flexibility; however, the results highlighted the importance of in-person therapy as well. Future research utilizing a greater number of participants and standardized measures of symptoms could be useful to determine whether teletherapy was genuinely an effective method in improving patients’ mental health over the course of treatment. Findings from this research could potentially advance online treatments for mental health disorders.

■ Acknowledgments

I gratefully acknowledge Ms. Hannah Dunn for extending to me a great amount of assistance.

■ References


■ Author

Brianna Son is a 12th grader at Flint Hill School in Virginia. She is planning to major in psychology and is interested in child developmental psychology.
Hierarchical Optimal Path Planning (HOPP) for Robotic Apple Harvesting

David W Liu
Athens Academy, 1281 Spartan Ln, Athens, GA, 30606, USA; david.weizhong.liu@gmail.com

ABSTRACT: Apples are among the most consumed fruits in the United States. Currently, almost all apples destined for the fresh market are picked by the human hand. Due to the shortage of seasonal manual labor and rising costs in manual apple picking, robotic apple harvesting has been explored for years. However, a challenge in developing and deploying an apple harvesting robotic system is how to deal with the unstructured apple tree environment. This work aims to demonstrate that structured 3D model representation of apple trees can significantly enable and facilitate robotic apple harvesting, particularly for optimal 3D path planning. Accordingly, a hierarchical optimal path planning (HOPP) algorithm is designed to significantly reduce or minimize the time cost during robotic apple harvesting in a 3D environment. The core idea of this HOPP algorithm is applying distance-constrained k-means clustering to group apples into 3D harvesting zones first, after which an optimal 3D robotic harvesting path is derived via the Traveling Salesman Problem (TSP) formulation and solution. Within each 3D apple harvesting zone, a second-stage optimal path planning is conducted by the TSP method on individual apples. Experiments showed that the proposed HOPP algorithm is promising.

KEYWORDS: Robotics; Agriculture; Apple Harvesting; Path Planning.

Introduction

Apples are America’s most consumed and favorite fruit and are either eaten raw or consumed in other ways. The American apple industry, with 5,000+ producers, brings billions of dollars to the economy each year and contributes nutritious fruit/food to consumers around the world. Remarkably, the most labor-intensive task in apple production is harvesting, and nearly all apples destined for the fresh market in America are picked by the human hand. Apple growers reported that harvesting labor accounts for approximately a third of their annual variable costs—such as much as pruning and thinning combined. In addition, manual apple harvesting is a time-sensitive operation where variable weather patterns generate uncertainty during employment planning and is costly. For example, the threat of an early fall frost could cause a short-term surge in the demand for apple pickers. Also, picking fresh market apples is both physically strenuous and highly repetitive. Apple picking exposes workers to fall hazards as well as ergonomic injuries through heavy lifting and repetitive hand actions. Therefore, to deal with the shortage of seasonal manual labor, rising costs, and risks to workers in manual apple picking, robotic apple harvesting has been explored for decades.

Robotics researchers have been actively working on the development of apple harvesting robots since the 1980s. In this field, visual fruit harvesting robots are common as reviewed recently. Typically, a fruit harvesting robot uses visual sensing, e.g., via 2D cameras, stereo vision systems, laser active vision systems, or multi-spectral imaging technologies, to perceive and learn fruit information such as location, shape, size, color, etc. The main technical tasks of those visual sensing systems include camera calibration, fruit target recognition and positioning, background recognition, 3D fruit reconstruction, and robotic path and trajectory planning. Those visual sensing systems also use visual servocontrol picking mechanisms to perform real fruit harvesting. However, a notable missing piece of those existing visual sensing systems for fruit harvesting is that they are not designed for global 3D mapping or reconstruction of all fruits on the entire tree, e.g., all apples on a tree. Instead, existing visual sensing techniques in robotic fruit harvesting are targeting individual or a few fruits in the local context. The lack of global information of distributions or localizations of all apples on a tree hampers the possibility of global optimal 3D path planning for actual apple harvesting. In response, this study contributes a unique investigation of how much gain can be achieved for optimal 3D path planning in robotic apple harvesting if 3D models of apples and their trees are available, which can stimulate and motivate future innovation in developing 3D reconstruction and mapping approaches for robotic apple harvesting and robotic fruit harvesting in general.

This work is also motivated by the fact that unstructured apple tree environment is a major challenge in developing and deploying apple harvesting robotic systems in real world scenarios, as pointed out and emphasized by literature papers. By leveraging an existing realistic 3D model of an apple tree and its hundreds of apples, this work will demonstrate that 3D representation of apples can significantly enable and facilitate robotic apple harvesting, particularly for optimal 3D path planning. Along this direction, a hierarchical optimal path planning (HOPP) algorithm is designed to significantly reduce or minimize the time cost during robotic apple harvesting in a 3D environment. Specifically, the 3D HOPP algorithm is composed of two stages: 1) applying distance-constrained k-means clustering to group apples into harvesting zones first, based on which an optimal
robotic harvesting path is derived via the Traveling Salesman Problem (TSP) formulation and solution; ² 2) within each apple harvesting zone, a second-stage optimal path planning is conducted by the TSP solution on individual apples. The above two stages of optimization form the HOPP algorithm in order to significantly reduce or minimize time cost in robotic apple harvesting. This work differentiates itself from other existing studies by contributing a new concept that 3D representation of apples can significantly enable and facilitate robotic apple harvesting and by contributing a novel HOPP algorithm.

## Methods

### 3D environment setup:

![3D environment setup](image)

**Figure 1:** Illustration of the 3D model of an apple tree and its apples. (A) Full 3D model of the apple tree. (B) Leaves are removed from the tree. (C) Small branches are removed further. (D) Trunk is removed further and only apples are kept.

This work uses a commercially purchased realistic 3D model of apple trees (from TurboSquid website) that includes 349 apples on it (represented in obj file format). Specifically, the apple tree's 3D model includes 640,768 vertices and 458,022 faces. Figure 1 illustrates several views of the apple tree and its apples. All of the HOPP algorithm developments and evaluations in this work are based on the 3D model in Figure 1. According to the 3D rendering of apples, it is observed that spatial distributions of apples are quite irregular, e.g., some regions have higher apple densities, while other regions have lower apple densities. Therefore, a straightforward intuition is that such irregular spatial distribution of apples should be leveraged for optimal 3D path planning in robotic apple harvesting, that is, the apple harvesting robot should give higher priority to those regions with higher apple densities and plan its 3D path accordingly.

### The HOPP algorithm:

Time cost is a key concern in robotic apple harvesting, e.g., the time costs of moving robotic shoulders and arms. In this work, I divided the total time cost of harvesting all apples into two stages: 1) moving the robotic shoulder across harvesting zones and 2) moving robotic arms to pick individual apples within each harvesting zone. Specifically, assume that the total number of robotic shoulder movements is \( k-1 \), that is, cost is shoulder. Then the total time cost of robotic shoulder movements is \( (k-1)t_{shoulder} \). Within the workspace of robotic arms (e.g., simplified as a 3D sphere centered at the joint of robotic shoulder and robotic arms) in each harvesting zone, the average time cost for picking an individual apple when only moving the robotic arms is represented by \( t_{arm} \). Then, to pick all the apples (n) on a tree, the time needed for individual apple harvesting when moving the robotic arms is \( n*t_{arm} \). Therefore, the total time cost of harvesting all apples, including both robotic arm and shoulder movements, will be \( (k-1)t_{shoulder} + n*t_{arm} \), which is the objective function of the HOPP algorithm to significantly reduce or minimize.

![Overall idea of the HOPP algorithm](image)

**Figure 2:** Overall idea of the HOPP algorithm. Here A-E represent five harvesting zones obtained by the distance-constrained k-means clustering. The path represented by orange arrows shows the first-stage optimal path across individual apples within a harvesting zone.

The overall idea of the 3D HOPP algorithm is simplified in a 2D space and illustrated in Figure 2. Specifically, in the first stage, all the apples will go through a distance-constrained k-means clustering so that these apples will be geographically grouped into neighboring harvesting zones (represented by the dashed green circles in Figure 2). Here, the harvesting robot arm’s workspace is simplified by a circle. In the 3D case, the robotic arms can reach anywhere in the 3D spherical volume in the harvesting zone without the need of moving the robotic shoulder. Then, the optimal robotic path connecting those harvesting zones is formulated as the TSP and is solved using heuristic greedy search solutions, as illustrated by the dashed orange arrows (A->B->C->D->E) in Figure 2. Similarly, in the 3D case, the optimal path is obtained by connecting those 3D spheres via the TSP solution. Intuitively, the step of distance-constrained k-means clustering can minimize the total number of apple harvesting zones. That is, given the fixed diameter of robotic arms’ spheric workspace, the distance-constrained k-means clustering can ensure that each grouped harvesting zone will include the largest possible number of apples, thus the total number of harvesting zones will be minimized. Also, the TSP solution to the cross-harvesting zone path can further minimize the average time cost of robotic shoulder movement. As a result, the total number of robotic shoulder movements k-1 and each shoulder movement’s average time cost is \( t_{shoulder} \).
are both minimized. Therefore, the total time cost of robotic shoulder movements \((k-1)t_{\text{tarm}} + n t_{\text{shoulder}}\) is minimized.

In the second stage, within each apple harvesting zone, an optimal path planning is further conducted on individual apples by the TSP formulation and solution again, as illustrated in the dashed blue circle and blue arrows in A' on the left side of Figure 2. Therefore, the average time cost \((t_{\text{tarm}})\) for picking an individual apple within each harvesting zone is minimized, and the total time needed for harvesting all individual apples \((n t_{\text{tarm}})\) when moving robotic arms is minimized. Thus, through the HOPP algorithm, the total time cost of harvesting all apples \((n (k-1)t_{\text{tarm}} + nt_{\text{tarm}})\), including time costs for both robotic arm and shoulder movements, is minimized.

**Distance-constrained k-means clustering for grouping apples into harvesting zones:**

Data clustering is a commonly employed methodology to group or partition a set of observational data (e.g., apples in this work) into collections with small within-group distances (e.g., within each apple harvesting zone) and big cross-group distances (e.g., across apple harvesting zones). Among various data clustering methodologies, one of the most popular methods is the k-means clustering algorithm. Briefly, the k-means clustering algorithm utilizes iterative refinement to generate a data clustering result after taking the number of clusters \(k\) and the dataset (e.g., apples’ spatial coordinates in this work) as input. The \(k\)-means clustering algorithm starts with an initial estimate for the \(k\) centroids, which can either be randomly created or randomly selected from the dataset. The algorithm then iterates between two steps of data assignment and centroid update. More specifically, in the data assignment step, each centroid defines one cluster, and each data point is assigned to its nearest centroid based on the metric of Euclidean distance, e.g., distance between apples’ spatial coordinates in this work. In the centroid update step, the cluster centroids are recomputed and updated by averaging the coordinates of all data points (e.g., apples’ spatial coordinates here) that are assigned to that centroid’s cluster. The \(k\)-means clustering algorithm iterates between the above two steps until a stopping criterion is met, i.e., no data points change clusters, the sum of the distances is minimized, or a maximum number of iterations is performed.

However, in this work’s context, directly applying the \(k\)-mean clustering algorithm to the apples’ data points would not work well due to the following reasons. First, it will be challenging to determine the right selection of cluster number \(k\) in our application scenario if the k-means clustering algorithm is directly applied. Second, it is hard to satisfy that the derived clusters will form valid apple harvesting zones, that is, all apples within a harvesting zone can be reached by the robotic arms, meaning that the distance between any apple and the centroid of the harvesting zone is shorter than the maximum length of the robotic arms. Therefore, a modified variant of \(k\)-mean clustering is adopted in this work’s context, that is, the distance-constrained \(k\)-mean clustering.

Mathematically, the following problem definition of distance-constrained \(k\)-mean clustering is adopted. Consider a collection of \(n\) apples’ 3D coordinate data, \(x_1,\ldots,x_n\), with \(x_i \in \mathbb{R}^3\), I aim to select a number of \(k\) groups or clusters (harvesting zones), \(C_1,\ldots,C_k\). It typically holds that \(k \ll n\). Each cluster \(C_j\) is assigned with a cluster centroid \(c_j \in \mathbb{R}^3\), which represents the center that apple cluster. An apple \(x_i \in \mathbb{R}^3\) is said to belong to cluster \(C_j\) if its distance from the centroid \(c_j\), with \(b_{ij}\), of every other cluster \(C_j\) is larger than its distance from \(C_j\). If \(x_i\) belongs to \(C_j\), I set a binary assignment variable \(r_{ij} \in \{0,1\}\) to 1, and to 0 otherwise. The solution of the apple clustering problem with distance constraint involves the computation of the optimal choice of cluster centroids \(c_j\), for \(j = 1,\ldots,k\), that minimizes the total of the distance of every apple \(x_i\) from the cluster it belongs to. More specifically, the problem is defined as follows.

I aim to find a choice of \(k\), of the cluster centroids, \(c_j \in \mathbb{R}^3\), and measurement data assignments, \(r_{ij} \in \{0,1\}\) that minimizes the objective function \(E\):

\[
E = \sum_i \sum_j r_{ij} \| x_i - c_j \|_2^2
\]

Subject to the following three constraints:

\[
\text{(I): } \sum_j r_{ij} = 1, \forall i = 1,\ldots,n
\]

\[
\text{(II): } \| x_i - c_j \|_2 \leq b_{ij}, \forall i = 1,\ldots,n, \forall j = 1,\ldots,k
\]

\[
\text{(III): } r_{ij} \in \{0,1\}, \forall i = 1,\ldots,n, \forall j = 1,\ldots,k
\]

The first constraint (I) along with the third one (III) implies that each apple is assigned exactly to one cluster (harvesting zone); the constraints (II) imply that \(b\) is the diameter of the harvesting zone and the distance between any apple in this harvesting zone and its centroid is less than or equal to its radius; finally, constraints (III) imply that \(r_{ij}\) are binary decision variables.

The definitions in Equations (1)-(4) fit our apple harvesting zone definition (illustrated in Figure 2) very well. Notably, in the literature, the constraint (II) in Equation (3) has not been considered, as far as I know. Including such a constraint, however, is quite useful in this work, since it will ensure the HOPP algorithm’s ability to define valid apple harvesting zones within predefined spherical robotic arm workspace. Meanwhile, the objective function in Equation (1) ensures that the total number of apple harvesting zones is minimized. To find a plausible solution to the problem defined in Equations (1)-(4), the method that used a Hegelsmann-Krause (HK) model was adopted in this work to find an appropriate \(k\) first before the distance-constrained \(k\)-means clustering. The core idea of the HK model is that data points with completely different attributes (e.g., geometric distances here) do not influence each other, while some sort of mediation occurs among data points whose distances are close enough. It has been justified and demonstrated that the HK model is well suited for the distance-constrained \(k\)-means clustering.

**Traveling salesman problem solution and hierarchical optimal 3D path planning:**

Once the apples have been clustered into harvesting zones via the methods in Equations (1)-(4), the spatial relationships among neighboring harvesting zones can be represented by a graph, e.g., spatially adjacent harvesting zones are linked by an edge, as illustrated in Figure 2. Then, the optimal 3D path planning at the first stage of harvesting zone is formulated as a TSP and solved using existing solution. Theoretically, TSP is
an NP-hard problem in combinatorial optimization and a heuristic greedy search algorithm is typically employed to obtain a near optimal result. Notably, the TSP formulation and solution has been utilized in many path planning problems in agriculture robotics and automation. Here, a heuristic greedy search implementation of TSP problem was adopted to find the optimal path given the clustered apple harvesting zones and their neighborhood graph, as illustrated in Figure 2. Thus, both the total number of robotic shoulder movements \( k \) and each shoulder movement’s average time cost is \( t_{\text{shoulder}} \) are minimized via the distance-constrained \( k\)-means clustering of apples and the TSP solution of optimal path planning at the harvesting zone level, respectively. Through this first stage of optimal 3D path planning, the total time cost of robotic shoulder movements \((k-1)t_{\text{shoulder}}\) is thus minimized.

Similarly, in the second stage of path planning, an optimal path search is further conducted on individual apples by the TSP formulation and solution again, as illustrated by the apples in A’ of Figure 2. As a result, the average time cost \( t_{\text{arm}} \) for picking an individual apple within each harvesting zone is minimized, and thus the total time cost for harvesting all individual apples \( (n^*t_{\text{arm}}) \) is minimized. Finally, the two-stage HOPP algorithm is summarized in Figure 3.

**Stage 1:**

- **Input:** \( n \) apples, represented by \( xi \) \((i=1,...,n)\)

- **Objective:**
  - Distance-constrained \( k\)-means clustering of \( xi \) \((i=1,...,n)\) into \( k \) apple harvesting zones.
  - The k-means clustering algorithm is used to determine the best \( k \) clusters.
  - The robot arm workspace’s sphere diameter \( \beta \) is used to constrain the size of each harvesting zone (Equation 2).

- **Output:** Hierarchical optimal 3D path across \( k \) harvesting zones.

**Stage 2:**

- **Input:** Optimal 3D path obtained by the TSP solution.

- **Objective:**
  - The derived optimal 3D path at the level of individual apples obtained by the TSP solution is represented by the white lines in Figure 4A and by the corresponding blue arrows Figure 4B. It is apparent in Figure 4A that each clustered harvesting zone exhibits high density of apples, meaning that each robotic shoulder movement can harvest as many apples as it can and therefore the time cost of robotic shoulder movement \((k-1)t_{\text{shoulder}}\) is minimized. Also, the optimal 3D path at the level of individual apples obtained by the TSP solution is represented by the white lines and shown in Figure 4C. In comparison, the path planning results by Method 2 are shown in Figures 4D-4F in a similar way. It is interesting that the planned path by Method 2 takes 40 robotic shoulder movements (for one random run) from one harvesting zone to another. That is, without taking advantage of global distribution of all apples and without the distance-constrained \( k\)-means clustering, the planned path by Method 2 takes 2.67 times the robotic shoulder movements (40 steps) to harvest all apples when compared to the HOPP method (15 steps). This result apparently suggests the effectiveness of the proposed HOPP method and the importance of taking the advantage of the global information of spatial distributions of all apples on trees for more efficient apple harvesting.

**Experiments:**

A series of experiments were designed and conducted to demonstrate the effectiveness and performance of the proposed HOPP algorithm summarized in Figure 3 and detailed in the above sections. The 3D models of apples in Figure 1 were used in these experiments.

As explained in Equation (3) and summarized in Figure 3, the parameter \( \beta \) describes the diameter of apple harvesting robot arms’ spherical workspace and it is a key factor that decides the output of the distance-constrained \( k\)-means clustering algorithm (Equations (1)-(4)). By alternating different \( \beta \) values, various apple harvesting zone clustering results were obtained by the HOPP algorithm. For quantitative comparison, a tabular path planning method (named Method 2 here) was considered as follows. It randomly selects apples to start with the harvesting process and uses greedy search in a local neighborhood for path planning. That is, Method 2 always goes from the current apple to the next neighbor with the shortest distance. Again, the average time cost for picking an individual apple within each harvesting zone is represented by \( t_{\text{arm}} \) and each shoulder movement’s average time cost is \( t_{\text{shoulder}} \). Then, I compare the total time costs for both path planning methods (HOPP and Method 2), that is, \((k-1)t_{\text{shoulder}} + n^*t_{\text{arm}}\). Here, it is assumed that the time cost \( t_{\text{shoulder}} \) is proportional to the distance between two harvesting zones (either minimized by the TSP optimization in HOPP or not in Method 2), and that the time cost \( t_{\text{arm}} \) is proportional to the distance between two individual apples (either minimized by the TSP optimization by HOPP or by greedy search in Method 2). It turned out that the time costs \( t_{\text{arm}} \) and \( n^*t_{\text{arm}} \) are quite close for HOPP and Method 2 as both methods employed a greedy search process among neighboring individual apples. Therefore, the following experimental comparisons focus on the time costs of robotic shoulder movements \((k-1)t_{\text{shoulder}}\).
Therefore, in terms of \((k-1) t_{\text{shoulder}}\), Method 2 takes 8.72 times the harvesting time by the HOPP method.

When \(\beta=2.2\), 9 apple harvesting zones were obtained by the HOPP method (Figures 6A-6B), and the optimal 3D optimal path is shown in Figure 6C. The path planning for Method 2 is shown in Figures 6D-6F. It is clear in Figure 6 that the proposed HOPP method takes significantly less time than Method 2. Again, Method 2 randomly selected the start apple to harvest and thus its total steps of robotic shoulder movements is random. I repeated the Method 2 for 30 times and the statistical results are shown in Figure 7. On average, the planned path by Method 2 takes 4.23 times the robotic shoulder movements (33.90 steps) to harvest all apples than that of the HOPP method (8 steps), as shown in Figure 7A. Also, the average path distances across harvesting zones by Method 2 (111.93) are 4.76 times more than those by the HOPP method (23.51), as shown in Figure 7B. Therefore, in terms of \((k-1) t_{\text{shoulder}}\), Method 2 takes 20.13 times more harvesting time than the HOPP method.

Reconstruction of 3D apple tree model:

Following the direction of 3D path planning for robotic apple harvesting and inspired by existing efforts of 3D reconstruction from image sequences, an initial effort was made to reconstruct 3D apple tree model in a lab environment. As shown in Figure 8, 63 photos (Figures 8A-8B) were taken around a laboratory apple tree by an iPhone from various angles and locations, and these photos were then used as the input for 3D reconstruction via the Meshroom software toolkit. The Live Reconstruction pane in Meshroom interface provides the visualization of each step in the 3D reconstruction, such as camera initialization and calibration, feature extraction, and structure from motion. Notably, the Meshroom toolkit utilizes a camera sensors database to determine the camera’s internal parameters and group them together. Therefore, Meshroom can infer the rigid scene structure (3D points) with the pose (position and orientation) and internal calibration of all cameras (iPhone camera’s different poses), as shown in Figure 8C. By using the calibrated cameras and the structure-from-motion step, Meshroom can generate a dense geometric surface of the 3D reconstruction of the apple tree, as shown in Figures 8C-8F. Although the 3D model’s quality can be improved in the future, the experiment in Figure 8 demonstrated the promise of reconstructing 3D models of apple trees using computer vision and computer graphics techniques. Given the rapid advancements of technologies, such as unmanned aerial vehicles (UAV) LiDAR or UAV camera data acquisition and 3D reconstruction, it is predicted that...
robotic apple harvesting guided by 3D path planning will become practical in the future.

![Figure 8: Reconstruction of 3D apple tree model. (A) Visualization of 63 images captured by iPhone camera from 63 poses. (B) Enlarged view of the first image in (A). (C) Reconstruction of the 3D apple tree model. (D)-(F): additional views of the 3D model in (C).](image)

Path planning for fruit harvesting robotics:

From a general perspective of robotic path planning, the HOPP algorithm proposed in this paper belongs to off-line path planning,¹² that is, the apple harvesting robot has prior access to complete information about the 3D environment of apples and their tree. Meanwhile, this HOPP algorithm can be applied online during harvesting if more information (e.g., apple localization and 3D reconstruction) becomes available. Also, the HOPP algorithm belongs to the category of point-to-point path planning, that is, the goal consists of determining a path from a starting point to a destination point by optimizing key parameters such as time and distance.¹² The HOPP algorithm is akin to the general category of cell decomposition method in the robotic path planning domain, which decomposes the free space into small regions called cells (e.g., apple harvesting zones in this work) and then searches for an optimal path in the cell graph using algorithms such as A*, Dijkstra, or TSP.¹² The HOPP algorithm also shares similarities with the category of coverage path planning (CPP) algorithms, in that CPP is defined as the task of determining a path that passes over all points of an area or volume (e.g., apple harvesting zones here) while avoiding obstacles.¹³ The HOPP algorithm and CPP share some common characteristics or requirements as the robot must cover the whole area (all apples), the robot should cover the entire region without overlapping, the robot should avoid all obstacles, the robot should use simple and smooth motion trajectories, and an optimal path is desired under considered criteria such as time and distance. However, it is not always possible to possess all these requirements or characteristics in complex agriculture environments like apple trees or orchards, and thus priority consideration is typically adopted. For instance, in its current form, the HOPP algorithm does not consider obstacle avoidance and does not plan robotic shoulder/arms motion trajectories.

There is much room to improve the HOPP algorithm in the future, such as considering obstacle avoidance, considering other optimization objectives like smoothness of robotic arms/shoulder movement, considering different time costs of robotic arm/shoulder movements, considering multiple simultaneous harvesting robots, and 3D path planning for harvesting many apple trees in the orchard. With these more complex considerations, the HOPP algorithm can be extended in various ways. For instances, the distance-constrained $k$-means clustering can be extended by adding additional obstacle avoidance into the optimization constraints, and the algorithm itself should be scaled to much larger number of apples like those in apple orchard. Also, the TSP formation and solution for finding the optimal path can be extended into a multiple TSP (mTSP) formation if there are several apple harvesting robots simultaneously participating in the apple picking.

It is known that the $k$-means clustering algorithm has a time complexity of $O(n^2)$, where $n$ is the input data size. This quadratic complexity debars the algorithm from being effectively used in large-scale applications. Researchers have explored lower-complexity implementation of the $k$-means clustering algorithm, such as the $O(n)$ complexity (linear order) counterpart of the $k$-means.¹⁴ In the future, such lower-complexity implementation could be employed into the HOPP algorithm for apple harvesting zone clustering via distance-constrained $k$-means clustering, which is the most time-consuming part of the 3D path planning algorithm. Also, such lower-complexity heuristic search algorithms can be explored for the TSP and mTSP solutions once the harvesting zones are mapped. These lower-complexity implementation of the 3D path planning algorithm is quite important for on-line path planning in large-scale apple harvesting in real-world orchards. Finally, this work does not provide a theoretical analysis of the algorithmic aspect of the HOPP method, which should be investigated in the future.

**Conclusion**

The world’s population is projected to reach nine billion people by the year 2050, which suggests that agricultural productivity must increase substantially and sustainably. The automatization of agricultural tasks, including fruit harvesting, is an essential step to deal with population growth. Various types of agricultural robots have been explored in the past few decades, however, how to deal with the unstructured agricultural environment is one of the hardest challenges. This work proposes a conceptually new HOPP algorithm for robotic apple harvesting by decomposing the challenging task into two-stage hierarchical apple harvesting zone clustering and optimal 3D path planning, and experimental results have demonstrated that the HOPP algorithm can gain significant benefits, e.g., minimizing apple harvesting time cost. In addition to the HOPP algorithm, another core contribution of this work is that it is demonstrated that the global information of spatial distributions of all apples should be leveraged for better apple harvesting. Although this proposed HOPP algorithm leverages the availability of 3D models of apples at current stage, it is expected that reconstructing 3D structures of apples and their trees from UAVs equipped with camera, LiDAR or multispectral imaging sensors and enabled with advanced AI tools will become practical in the future. Also, despite that this work simplifies the apple tree environment and harvesting execution, the HOPP algorithm itself is scalable and extendable.
by including a variety of other considerations, such as obstacle avoidance and multiple harvesting robots, in the future.

**Acknowledgements**

I would like to thank Professor Charlie Li, Dr. Javier Rodriguez and Daniel Petti of Bio-Sensing and Instrumentation Lab, Phenomics and Plant Robotics Center, College of Engineering, University of Georgia, for their guidance and help when I performed an internship in their lab in Summer 2021 and Fall 2021. I would like to thank Zhengliang Liu and Jiaying Gao for their guidance and help on data visualization and Python programming. I would like to thank my teacher Mike Callinan at Athens Academy who has been teaching me robotics courses and guiding me on robotics projects for over four years. Finally, I would like to thank my father for his guidance and support during my research.

**References**


**Authors**

David W. Liu is a 10th grader in Athens Academy, Athens, GA. He is interested in robotics, computer programming, and mathematics. He has been taking robotics courses in Athens Academy since 6th grade and has been actively involved in multiple robotics activities such as FLL and FTC challenges since 6th grade.
The Rationality Analysis and Prediction of THE World University Rankings

Yingao Gao
Mount Douglas Secondary School, 3970 Gordon Head Rd, Victoria, BC, Canada; gcliclig@gmail.com

ABSTRACT: In the perspective of the wide population of high-school teens, choosing a suitable university can help students take advantage of their strengths better, and ultimately affect their life journey. In general, the world university rankings analyze each post-secondary college through the aspects of their teaching, research, and the future employment of their students, which provides important information for students and guides their choices. However, different rankings have distinct focuses and different weights on each indicator, which may mislead students, and possibly cause confusion. Therefore, this paper takes THE Rankings as an example, utilizes statistical theories to analyze the rationality of performance indicators, and uses principal component analysis (PCA) to optimize the weights of each indicator, so the scores and rankings will accurately reflect the characteristics, strengths and weaknesses of each university. Also, the data of the top 200 universities in the past years is used to train a neural network model to predict the future rankings, providing a reference for senior high-school students.

KEYWORDS: Mathematics; Analysis; THE World University Rankings; Principal Component Analysis; Neural Network.

Introduction

THE Rankings is a world university ranking published by Times Higher Education, which judges the universities on 5 indicators, including Teaching, Research, Citations, International Outlook, and Industry Income. These 5 indicators are then split into 13 more precise indicators, as shown in Table 1. Until 2021, THE Rankings covered more than 1500 universities in 93 countries and territories.

Table 1 shows that reputation impacts the scores of indicators heavily, which affects Teaching and Research the most. Although reputation can reflect the level of universities, it is hard to measure accurately and fairly, and is mainly subjective. The weights of indicators reflect the focus of the ranking, but not all of the indicators are fairly weighted. Bowman et al. propose that the rankings are affected by the anchoring effect which means that they may be swayed by other institutions or rankings, and the authors also provide plans to optimize the reputations investigating method.³ Others believe that among the perfectly seeming grading system, there are only a small number of indicators that contribute heavily to the final ranking, which can cause fraud.⁴ Therefore, our goal is to optimize the indicators so that they can accurately and truthfully evaluate universities to provide students with relatively reasonable references.

The world university rankings have strong hysteresis. When universities implement new policies on teaching or research, it often takes years to take effect. The Universities’ reputations get affected by many different factors, and it is very hard to change the stereotype of people. Therefore, universities cannot maintain the current or future standard as the same as the one used in the ranking. Ranking predictions based on machine learning can utilize a massive amount of data from past years and seek out patterns, coming up with relatively reliable results, which solves the hysteresis problem to some extent, and helps high-school students to find suitable post-secondary education opportunities.

This paper uses data from the official website of THE rankings and adopts a data fitting method and a correlation method to analyze the data from both statistical and graphical perspectives, to optimize the existing ranking system. Also, this paper trains a back propagation neural network using past years’ data to predict upcoming ranking.

Methods

Indicator Analysis:

To optimize the weights of indicators, a deep understanding on the rationality of the current indicators is needed. This paper determines the rationality of indicators in two aspects: First, the score of an indicator should correlate closely with the rank. That is, the higher a university scores on an indicator, in
general, the higher its rank is. Besides, the score distribution of the indicators should be reasonable, so it can differentiate the different levels of universities.

Least square fitting⁵ is used to determine whether there is a correlation between each indicator and ranking. Assume X = (x₁, x₂, ..., xᵢ, ..., xᵣ) is a set with N data points, where x is the ranking, and Y is the score. We find a straight-line \( l: y = kx + b \), where the sum of the distance of every point \((xᵢ, yᵢ)\) in \( A \) to the corresponding point \((kxᵢ, Yᵢ)\) on line \( l \) is the smallest, which can be expressed as Equation (1).

\[
L = \sum_{i=1}^{N} |yᵢ - kxᵢ + b|
\]

(1)

The distance \( L \) is determined by the parameters \( k \) and \( b \), which can be solved using the polyfit() function in MATLAB.

Although least square fitting can show the correlation between scores and ranking intuitively, it cannot quantitatively analyze the correlation. Calculating the correlation coefficient of the ranking and score is one solution to evaluate the correlation. Pearson coefficient is used here and the larger absolute value of it indicates the higher correlation.⁶ Positive or negative indicates whether the two are positively correlated or negatively correlated. The Pearson coefficient can be calculated by Equation (2).

\[
P = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}
\]

(2)

where \( \text{cov()} \) is the covariance, \( \sigma \) is the standard deviation, \( X \) and \( Y \) represent the score of an indicator and the ranking respectively. The covariance is used to measure the variation trend of \( X \) and \( Y \). If the covariance is negative, it indicates that the overall trend of the two is opposite and negatively correlated; if the covariance is positive, the two are positively correlated. The standard deviation in the denominator is to normalize the covariance to ensure that it can accurately reflect the correlation between \( X \) and \( Y \).

Indicator Optimization:

The previous analysis shows that some indicators have problems in weight allocation or scoring standards. Some indicators have no contribution to the results or have no negative effect on the evaluation of universities, so that the ranking results cannot fully represent the actual level of universities. Since the scores of each indicator cannot be modified, the strategy of weight optimization is adopted here to modify the contribution of each indicator to the ranking, so as to reduce the influence of unreasonable indicators and scores on the ranking. Principal component analysis (PCA) is used here to find a better weight for each indicator.⁷

PCA is a statistical method that converts a group of variables that may be correlated into linearly independent variables (principal components) through orthogonal transformation, and utilizes the original information contained in the principal components to conduct subject analysis. To determine the weight, the principal components should be expressed as the linear combination of the original indicators, following Equation (3).

\[
F_i = a_1X_1 + a_2X_2 + a_3X_3 + a_4X_4 + a_5X_5
\]

(3)

where \( X \) is the standardized score of each indicator, and \( a \) is the eigenvector corresponding to the eigenvalue of the covariance matrix of \( X \).

Ranking Prediction:

The world university ranking is usually hysteretic, with changes in teaching policies often not reflected in the rankings until several years later. The ranking of a certain university may change greatly over time. For example, Peking University ranked 42nd in 2016, 30th in 2019, and 24th in 2020.⁸ In order to better reflect the current level of universities and provide some references for students who are about to choose universities, this paper tries to predict the rankings.

It is difficult to predict the ranking directly since it is influenced by scores of various indicators and other external factors. Therefore, this paper only predicts the total score (the score of an indicator is also possible) of a certain university. In the later stage, the total score can be used to calculate the rankings. The mean square error (MSE) is used to measure the difference between the predicted scores and the ground truth to verify the feasibility of the proposed method, which is given by Equation (5).

\[
\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2
\]

(5)

where \( Y \) is the predicted scores and \( \hat{Y} \) is the ground truth. It is generally believed that the method is feasible when the difference between the predicted score and the ground truth is within 5 points, that is, the MSE is less than 25.

In this paper, a back propagation neural network is proposed to predict scores.⁹ The neural network imitates the connection between nerve cells of the human brain to transmit information. The main idea is to use a large amount of data to train the model, so as to complete corresponding tasks under similar conditions.¹⁰ In the training phase, the loss function is utilized to judge the error between the output and the ground truth, and model parameters are updated by a back propagation algorithm to make the output of the next epoch closer to the ground truth. By comparing the performance of models with different layers, this paper chooses to build a neural network model with five hidden layers, as shown in Figure 1.

![Figure 1: The score prediction neural network.](image)

Results and Discussion

Indicator Analysis:

This paper uses polyfit() function to fit and analyze the relationship between the scores of each indicators and the ranking. Figure 2 shows the least square fitting results of the top 200 universities in 2020 THE Rankings, which shows the correlation between ranking and the score of each indicator intuitively.
Also, the number of moderate level universities should be that fall in the interval of high or low scores should be less. The distribution in a certain range, and the number of universities schools, namely the scores should not be too concentrated in ranking, which is consistent with the conclusion of least national Outlook and Industry Income have a low correlation Research has the highest correlation with ranking, while Inter.

As shown in Figure 2, Teaching, Research, and Citations are strongly correlated with ranking, while International Outlook and Industry Income are weak. The analysis shows that the scores of International Outlook and Industry Income have smaller weights and have less impact on ranking than the other three indicators. In addition, it also reflects the problems in the correlation among indicators, that is, Teaching, Research, and Citations are more correlated. The above conclusions are consistent with the weight distribution in Table 1. However, some other different conclusions can be drawn in Figure 2. The weights of teaching, research, and the citations are the same, but the sum of the squared residuals of the citations is significantly larger than the first two, which means there may be some problems with the scoring of the citations. Besides, although the weights of the international outlook and industry income are relatively low, it can be seen from the data points that the scores of the two indicators have almost no effect on the ranking, whether the two indicators can be reflected in the final ranking is worth considering.

Then, I made a quantitative analysis of the correlation between each indicator and the ranking. In MATLAB, we can directly calculate Pearson correlation coefficients by using corr() function. The results of coefficients between scores and rankings are shown in Table 2.

**Table 2:** Pearson coefficient between indicators and rankings. The larger the absolute value of the coefficient, the greater the correlation with the ranking.

<table>
<thead>
<tr>
<th>Indicators</th>
<th>Teaching</th>
<th>Research</th>
<th>Citation</th>
<th>International Outlook</th>
<th>Industry Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>-0.8096</td>
<td>-0.8560</td>
<td>-0.4096</td>
<td>-0.1082</td>
<td>-0.0907</td>
</tr>
</tbody>
</table>

Table 2 quantitatively shows that the score of Teaching and Research has the highest correlation with ranking, while International Outlook and Industry Income have a low correlation with ranking, which is consistent with the conclusion of least square fitting.

Next, I evaluated the rationality of scoring. Rational distribution of scores should be able to separate the good and poor schools, namely the scores should not be too concentrated in distribution in a certain range, and the number of universities that fall in the interval of high or low scores should be less. Also, the number of the moderate level universities should be relatively greater, which is in the common cognition of the evaluation of things. The data of the top 200 universities in THE Rankings in 2020 is used here to draw a score distribution histogram to analyze the rationality of indicators, as shown in Figure 3.

**Figure 3:** Histogram of score distribution of five indicators. There may be some problems with the Citations score distribution.

The score distribution of Teaching and Research is relatively reasonable, while there are many extreme scores in Citations and Industry Income. Taking Citations as an example, more than half of the universities scored more than 80 points, which is equivalent to adding roughly the same score to most of the universities. Besides, the citations of a paper depend on a variety of factors, including academic level, publication date, and field of study. Because of all these factors, it has little impact on the overall rankings, and it is difficult to distinguish the quality of the published paper. Theoretically, the weight of this kind of indicator should be reduced accordingly.

By analyzing the indicators and weights of THE Rankings, we can see that there are indeed some problems in the evaluation system that affect the reliability of rankings. The weak correlation between some indicators and the ranking or no correlation between indicators reflect the unreasonable selection of indicators and weight allocation in the ranking, which need to be specially adjusted for these problems.

**Indicator Optimization:**

The pca() function in MATLAB can be used to obtain the information content E in the five principal components. The results are shown in Table 3. The first four principal components already contain more than 98% of the information, so only the first four are used for subsequent analyses.

**Table 3:** The information content of each principal component.

<table>
<thead>
<tr>
<th>Principal Component</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
<th>F5</th>
</tr>
</thead>
</table>

The relative weights $W$ of indicators in Equation (4) can be obtained by evaluating the importance of each principal component with the amount of information it contained. The final weight is the normalized weight vector $W$, as shown in Table 4.

**Table 4:** The comparison of weights before and after optimization.

<table>
<thead>
<tr>
<th>Indicators</th>
<th>Teaching</th>
<th>Research</th>
<th>Citations</th>
<th>International Outlook</th>
<th>Industry Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Weights</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.075</td>
<td>0.025</td>
</tr>
<tr>
<td>Optimized Weights</td>
<td>0.2203</td>
<td>0.2714</td>
<td>0.2136</td>
<td>0.2011</td>
<td>0.0903</td>
</tr>
</tbody>
</table>
The optimization results show that the weights of Teaching, Research, and Citations have decreased to varying degrees, while the weights of International Outlook and Industry Income have increased. As previously analyzed, the correlation between rankings and International Outlook or Industry Income is weak, because the original weights of the two indexes are too small to be reflected in the final rankings. The weights of Teaching and Research decrease since both of them are affected by reputation, and the decrease in the weight is conducive to improving the objectivity of the ranking. The decrease in Citations may be due to the unreasonable scoring range of this indicator, which needs to reduce its influence on ranking.

The results of the optimization and analyses methods in this paper verify each other, which proves that the analyses of indicator and weight optimization are reasonable to a certain extent. The method proposed in this paper is not the only and not necessarily the optimal optimization method. The purpose is to provide a new way to analyze the reliability of world university rankings by using the principles of mathematics and statistics.

**Ranking Prediction:**

I collected the scores of the top 200 universities in THE Rankings from 2011 to 2016 and selected 147 of them that have been in the top 200 for six years. The 147 universities were divided into three parts: a training set, validation set, and test set, with a data ratio of 80:20:47. They were used to learn data features, update parameters, and test performance respectively. With the scores from 2011 to 2015 as input, the neural network model learned the features and changed the trend of scores in the five years and predicted the scores in 2016. Finally, the MSE of the neural network reaches 12.21, which means that the average error between the ground truth and the predicted value is only about 3.5 points, which proves that it is relatively reliable to use neural network to predict scores.

**Conclusion**

This paper analyzes the rationality of THE World University Rankings. The correlation between indicator scores and rankings are used to measure the rationality of the indicators, using the least square fitting and the Pearson correlation coefficient to filter out the unreasonable indicators from both statistical and graphical perspectives. The Principal Component Analysis is used to optimize the weights of indicators and combines the optimized results with theoretical analysis for validation. In addition, this paper designs a back propagation neural network model, which contains 5 hidden layers, using data from years 2011-2016 to train the model, and data from 2017 to check the model. It has an error margin of around 3.5, so it can predict future rankings relatively accurately.

The purpose of this paper is to provide new ideas and methods for the analysis of the world university rankings, hoping to offer an improved reference for many high school students to choose their own university. Due to my deficiencies in the breadth and depth of relevant knowledge, this paper still has deficiencies in the rationality evaluation criteria of weights and the robustness optimization of neural network performance, and I still need further study and research in this.

**Acknowledgments**

I would first like to express my highest gratitude to my parents for their endless help in my research. In addition, I would like to thank Dr. Chang for providing me with the ranking data of the THE World University Rankings. Finally, I am thankful to the IJHSR team for providing the platform to show my research.

**References**


**Author**

Yinggao Gao is currently a junior at Mount Douglas Secondary in Victoria, BC. He gained interest in math and computer science by participating in competitions and hopes to continue studying these in post-secondary school.
Simulation of Heat Diffusion in a Tidally-Locked Exoplanet

Giulia Murgia
Istituto di Istruzione Superiore G. Asproni, Loc. Su Pardu, Iglesias, I-09016, Italy; g.murgia03@gmail.com

ABSTRACT: I perform a numerical calculation of the heat diffusion in a simulated tidally-locked exoplanet. I solve the heat diffusion equation for the surface and the interior of the planet using a finite element grid. I develop a C++ code specifically for this purpose. I take into account both the energy received from the star, the energy radiated into space, the heat source provided by the radioactive decay inside the planet, and the effects of a single-layer atmosphere model. I apply the simulation to the case of Proxima Centauri b. The evolution of the temperature distribution is obtained in the time range from 0 to 15 Gyr. The most important results are the prediction of the possible formation of underground oceans in the dark side and the identification of the habitable zone on the surface of the exoplanet.

KEYWORDS: Physics and Astronomy; Computational Physics; Exoplanets; C++ Simulation; Heat Diffusion.

Introduction

There is a growing number of exoplanets that are being discovered. Many of them are rocky planets, similar to Earth, that orbit in the habitable zone. This is a region around a star in which an exoplanet can host liquid water on its surface and possibly support life.¹,² The presence of liquid water is possible only at a narrow range of temperatures. On the Earth’s surface, at atmospheric pressure of 10⁵ Pa, this range is from about 270 to 370 K, but it may vary if the pressure of the exoplanet’s atmosphere is significantly different.³ Understanding the thermal evolution of an exoplanet is fundamental to determining where and when the conditions for habitable zones are satisfied.

The temperature reached by a planet during its evolution depends on several factors: it absorbs a part of the electromagnetic radiation emitted by its star; it produces internal heat with radioactive decay in the planet’s core; it diffuses a part of its heat into space, according to the Stefan-Boltzmann law.

A critical parameter is the planetary equilibrium temperature, which is the theoretical temperature that a planet would reach if it were a black body. Thermal equilibrium induced by radiation is reached when the power transmitted by the star to the planet will be equal to that lost by the planet in space. The temperature, $T_{eq}$, at which this equilibrium is reached, averaged over the planet’s surface, is equal to:

$$T_{eq} = \left( \frac{L_*}{4 \pi D^2 \sigma} \right)^{1/4}$$

where $a$ is the albedo, $L_*$ is the star luminosity, $\sigma$ is the Stefan-Boltzmann constant, and $D$ is the orbit semi-axis.⁴ The albedo is the fraction of the star radiation reflected into space by clouds, sea-ice, and other bodies on the planet’s surface.

The planet’s atmosphere plays an essential role for two reasons. First, it contributes to raising the average temperature due to the greenhouse effect. Second, it contributes with the winds to transport heat between the illuminated face and the dark one, partly rebalancing the distribution of surface temperature.⁵

I considered a simplified scenario of a single-layer atmosphere model⁶ that I slightly altered to calculate the total energy budget as shown in Figure 1. At the equilibrium:

$$\sigma T_{eq}^4 + \epsilon_P T_P^4 = \epsilon_P T_P^4 + 2 \epsilon_{atm} \sigma T_{atm}^4$$

where the planet and the atmosphere are supposed to be gray bodies with emissivities of $\epsilon_P$ and $\epsilon_{atm}$, respectively.

Thus, the temperatures of the atmosphere and the planet are expected to be

$$T_{atm} = [2 - \epsilon_{atm}]^{-1/4} T_{eq}$$

$$T_P = \left( \frac{\epsilon_P}{2 - \epsilon_{atm}} \right)^{1/4} T_{eq}.$$ 

In fact, from the second equation in Eq. 2, it follows that $\epsilon_P T_P^4 = 2 T_{atm}^4$. Substituting this relation in the first equation, it results in $T_{atm} = (2 - \epsilon_{atm}) T_{eq}^4$, from which the first of Eq. 3 is derived. By eliminating $T_{atm}^4$ from the first equation in Eq. 2 and rearranging the terms, it follows $T_{eq}^4 (1 + \epsilon_{atm} (2 - \epsilon_{atm})) = \epsilon_P T_P^4 / (\epsilon_P (2 \epsilon_{atm}))$, from which the second equation in Eq. 3 is obtained.

Note that $T_{eq}$, as defined in Eq. 1, appears in Eq. 3 only as a reference for the average incoming stellar radiation flux, but in general neither the atmosphere nor the planet’s surface is
at a temperature $T_{eq}$ at the equilibrium. In the limit $\epsilon_{atm} \to 1$, the atmosphere temperature could approach the equilibrium temperature $T_{eq}$. However, because of the greenhouse effect, the temperature of the planet's surface $T_P$ remains always greater than $T_{eq}$ if $\epsilon_{atm} > 0$, even if $\epsilon_{atm} \to 1$.

In this paper, I present a simulation of the heat diffusion for the exoplanet Proxima Centauri b. It orbits around Proxima Centauri, a red dwarf star with a mass of about 12.5% of the Sun’s mass and luminosity of about only 0.1% that of the Sun. Proxima Centauri b is an exoplanet very similar to Earth located at only 0.0485 UA. Proxima Centauri b and its star form an isolated star-planet system subjected only to gravitational tides. Therefore, there is an high possibility that Proxima Centauri b is tidally locked in a nearly circular orbit.⁷

Proxima Centauri b is located in the habitable zone⁸ and, considering an albedo of 0.3, it has an equilibrium temperature of 229 K. Assuming a planet emissivity of 0.9, and an atmosphere emissivity of 0.9, the temperature of the planet is 273 K and the temperature of the atmosphere is 223 K.

**Methods**

*Spherical coordinates:*

Starting from Fourier’s law of heat conduction, I obtained the heat diffusion equation in spherical coordinates⁹:

$$ q = -k \frac{dT}{dr}, $$(4)

where $q$ is the heat flux in [W] through the surface $A$ in [m²], $k$ is the material conductivity in [W/(m·K)], and $d T/d L$ is the temperature gradient perpendicular to the surface.

I chose to represent the temperature on a grid of nodes in spherical coordinates. Using Eq. 4, I calculated the heat flux through the faces of an infinitesimal cell centered in $r, \phi, \theta$, (see Figure 2) obtaining the following

$$ q_r = -k r^2 \frac{d \phi}{sin \theta} \frac{dT}{dr}, $$

$$ q_{\phi} = -k \frac{dr}{sin \theta} \frac{dT}{d \phi}, $$

$$ q_{\theta} = -k r dr \frac{dT}{d \theta}. $$

(5)

![Figure 2:](image)

Considering a heat capacity $c_P$ and a density $\rho$, from

$$ dQ = c_P \rho m dT $$

Where $dQ$ is the amount of heat that must be added (at constant pressure) to an object of mass $m$ to raise its temperature by $dT$, I obtained

$$ q = c_P \rho V \frac{dT}{dt}, $$

(7)

where $V = r^2 \sin \theta d\phi d\theta$.

Developing in the Taylor series $q_{r+dr}$ up to the first order

$$ q_{r+dr} = q_r + \frac{\partial q_r}{\partial r} dr $$

(8)

The net heat flux in the cell is:

$$ q_{cell} = q_{r+dr} - q_r = \frac{\partial q_r}{\partial r} dr. $$

(9)

By repeating for all directions and by adding the three contributions, using Eq. 5 and Eq. 7, I solved for the heat diffusion equation in spherical coordinates

$$ \frac{\partial T}{\partial t} = \alpha \left[ \frac{\partial^2 T}{\partial r^2} + \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \phi} \left( \sin \theta \frac{\partial T}{\partial \phi} \right) \right] + Q_{\text{decay}}, $$

(10)

where $\alpha = k c_P \rho$ is the diffusivity constant in [m²/s] and $Q_{\text{decay}}$ is the power generated per unit volume by the internal nuclear decays in [W/m³].¹⁰

I further developed the derivatives of the product in $r$ and $\theta$ re-writing:

$$ \frac{\partial}{\partial r} \left[ \frac{\partial T}{\partial r} \right] \frac{\partial^2 T}{\partial r^2} + \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \phi} \left( \sin \theta \frac{\partial T}{\partial \phi} \right) \frac{\partial^2 T}{\partial \phi^2} + \frac{\partial T}{\partial \theta} \frac{\partial^2 T}{\partial \theta^2} = 0. $$

(11)

and

$$ \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial T}{\partial \theta} \right) \frac{\partial^2 T}{\partial \theta^2} + \sin \theta \frac{\partial T}{\partial \phi} \frac{\partial^2 T}{\partial \phi^2} + \cos \theta \frac{\partial^2 T}{\partial \phi \partial \theta} = 0. $$

(12)

I solved Eq. 10 using a finite element grid. I developed a dedicated C++ code specifically for this purpose.

Each cell of the grid is centered in a node at position $(r, \phi, \theta)$ and has a volume of $\Delta V = r \Delta r \cdot r^2 \sin \theta \Delta \phi \cdot r \Delta \theta$. Indeed, to solve Eq. 10 it is necessary to have both first and second derivatives. Therefore, I developed the temperature in the Taylor series up to the second order. For example, let’s consider the radial direction

$$ T(r+\Delta r) = T(r) + \frac{\partial T}{\partial r} \Delta r + \frac{\partial^2 T}{\partial r^2} \frac{\Delta r^2}{2} + \frac{\partial^3 T}{\partial r^3} \frac{\Delta r^3}{6} + \cdots. $$

(13)

Adding and subtracting, I found the first and second derivatives

$$ \frac{\partial T}{\partial r} = \frac{T(r+\Delta r) - T(r-\Delta r)}{2 \Delta r}, $$

$$ \frac{\partial^2 T}{\partial r^2} = \frac{T(r+\Delta r) - 2 T(r) + T(r-\Delta r)}{\Delta r^2}. $$

(14)

These formulas allow me to compute the derivatives of the temperature for each node of the cell considering the temperatures of the node before and of the node after, along each direction.
Time Evolution:

The evolution of temperature is obtained by discretizing the time in steps of finite duration $\Delta t$. I first calculated the spatial derivatives at a given time and then I use Euler’s formula to extrapolate the temperature at the next time step

$$T(t+\Delta t) = T(t) + \frac{\partial T}{\partial t} \cdot \Delta t + \ldots$$  \hspace{1cm} (15)

where the time derivative at a time $t$ is provided by Eq. 10.

Euler’s method is known to be stable, so the time step $\Delta t$ is chosen adaptively such that at each time the maximum variation of temperature is 10%.

One-Dimension test:

I tested the code in the one-dimensional case along the planet’s equator. Only heat diffusion was taken into account.

The initial condition is $T_0 = (300 - 20\sin \phi) K$. I used a periodic boundary condition since the equator is a closed ring of radius $r$. In this simple situation, it is possible to derive an analytical solution to Eq. 10: $T(\phi, t) = (300 - 20\sin \phi) e^{-t/\tau_c} K$, where the characteristic time is given by $\tau_c = r^2/\alpha$. By assuming $\alpha = 2 m^2/s$, $d\phi = 0.36^\circ$, and $r = 1000 K m$, $\tau_c = 15 Gyr$.

In Figure 3 I show the comparison of the numerical results (red dots) with the analytical solution (blue lines).

Three-Dimension Simulation:

I applied the numerical method described in the section above to the case of the exoplanet Proxima Centauri b. I performed a three-dimension simulation adopting the physical parameters listed in Table 1.

For decay power, density, thermal conductivity, and specific heat capacity I took Earth’s average values. For simplicity, I assumed that all physical parameters are uniform in space and constant in time, except for the radioactive decay that decreases with time according to

$$Q_{\text{decay}} = 0.01 e^{-t/\tau_{\text{decay}}}[\mu W/m^3].$$  \hspace{1cm} (16)

The decay of radioactive isotopes like uranium-235 (235U) contributed a large fraction of radiogenic heat to the early Earth due to its relatively short half-life, $\tau_d$. Indeed, for this simulation, I assumed $t_{\text{decay}} = 1 Gyr$ by considering that for 235U $\tau_d = 0.7 Gyr$ and that $t_{\text{decay}} = 2t_d ln(2)$.

In addition, I considered that the radioactive decay is uniformly distributed in the planet’s interior, with no radial dependence. No radioactive decay is considered to be present in the atmosphere.

For the Proxima Centauri planetary system, I considered the parameter values reported in the NASA Exoplanet Archive.

It should be considered that some of the fiducial model parameters used to run the simulation are more uncertain than others. Parameters like, the stellar luminosity, the distance of the planet from the star, and the planet radius are reliable estimates since they came from astronomical observations of the Proxima Centauri system. Others, like the thermal conductivity, are more uncertain because they depend on the composition of the interior of the exoplanet and are not directly accessible. In these cases, average Earth values are assumed. Furthermore, the assumption of constant composition throughout the entire radius of the planet should be considered as a zeroth-order approximation.

The C++ program used for the simulation is named PLANET CODE and is available in GitHub: https://github.com/GiuliaMurgia03/Simulation-of-Heat-Diffusion-in-a-Tidally-Locked-Exoplanet.

Table 1: Simulation set-up.

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho_{\text{crust}}$</td>
<td>5558 Kg/m$^3$</td>
</tr>
<tr>
<td>$k$</td>
<td>5 W/(K m)</td>
</tr>
<tr>
<td>Initial $Q_{\text{decay}}$</td>
<td>0.01 $\mu W/m^3$</td>
</tr>
<tr>
<td>$t_{\text{decay}}$</td>
<td>1 Gyr</td>
</tr>
<tr>
<td>$c_p$</td>
<td>200 J/(kg K)</td>
</tr>
<tr>
<td>$\epsilon_p$</td>
<td>0.9</td>
</tr>
<tr>
<td>$\epsilon_{\text{atmosphere}}$</td>
<td>0.9</td>
</tr>
<tr>
<td>$L$</td>
<td>0.001154 $L_0$</td>
</tr>
<tr>
<td>$r$</td>
<td>6680 Km</td>
</tr>
<tr>
<td>$\delta$</td>
<td>0.0485 $L_A$</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.3</td>
</tr>
<tr>
<td>$\Delta r$</td>
<td>430 Km</td>
</tr>
<tr>
<td>$\Delta \phi$</td>
<td>5.6 deg</td>
</tr>
<tr>
<td>$\Delta \theta$</td>
<td>5.6 deg</td>
</tr>
<tr>
<td>$T_{\text{eq}}$</td>
<td>229 K</td>
</tr>
<tr>
<td>Number of grid nodes</td>
<td>16x64x32</td>
</tr>
<tr>
<td>$T_{\text{eq}}$ thermal equilibrium temperature</td>
<td>229 K</td>
</tr>
</tbody>
</table>

Initial conditions:

The simulation begins when the planet has already finished the accretion phase by reaching the final radius. I assumed that the planet has been heated by the release of gravitational energy to a uniform temperature of 1000 K.

Boundary conditions:

I used a periodic boundary condition along the longitudinal direction ($\phi$). At the north and south poles of the grid, I assumed that there is no heat flux for $\theta = 0$ deg and $\theta = 180$ deg along the latitude direction. Thus, according to Eq. 4, $\partial T/\partial \theta = 0$. Indeed, from Eq. 14 I obtain $T(\theta + \Delta \theta) = T(\theta - \Delta \theta)$. I applied the same condition along the radial direction for $r < 0$ at the center of the planet, obtaining $T(r - \Delta r) = T(r + \Delta r)$.

For the illuminated side, I considered a radiative boundary condition in which

$$-k \frac{\partial T}{\partial r} = A \alpha \epsilon_p T^5 - 4 A \alpha \epsilon_p \sin \theta \cos \phi -$$

$$- \epsilon_{\text{atm}} A \alpha T_{\text{atm}}.$$

I obtained

$$\frac{\partial T}{\partial r} = \frac{\epsilon_p}{k} (4 \cos \phi \sin \theta T_{\phi \theta}^5 +$$

$$+ \epsilon_{\text{atm}} T_{\text{atm}} - \epsilon_p T^5),$$

where $T_{\phi \theta}$ is the temperature at the equator.

DOI: 10.36838/v4i3.4
from which I got the radiative boundary condition
\[ T(r+\Delta r) = T(r-\Delta r) + \frac{2\Delta r \sigma}{k} \left[ \frac{4\cos \phi \sin \theta T_\text{a}^4}{e_{\text{an}} T_{\text{e}4} - e_T T^4} \right]. \] (19)

For the dark side, only the radiation from the atmosphere contributes to the heat flux toward the surface and the second term on the right-hand side of Eq.17 is null.

\section{Results}
I presented the evolution of the temperature for Proxima Centauri b in the time range from 0 to 15 Gyr in Figure 4. The simulation run took about a day of computing time on my laptop. The images were obtained with PLANET CODE using a visualization program written in C++ and OpenGL.\textsuperscript{14} I show the northern hemisphere of the illuminated side of the planet, with a sector removed to illustrate the interior.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure4.png}
\caption{Each cell of the grid is colored based on its temperature as represented in the color bar.}
\end{figure}

At the beginning (t=0 Gyr), the planet is at an average temperature of 1000 K and the surface is starting to cool down due to the radiative losses. After 0.5 Gyr the surface already reaches its equilibrium temperature, while the interior is warming up due to the heat released by nuclear decay.

Note that at the center of the illuminated side, the temperature is between 313-333 K. Around this central hot area, at mid-latitudes, a ring forms with a temperature between 273 and 313 K. Assuming for the exoplanet an atmospheric pressure of 10^5 Pa, these temperatures permit the existence of liquid water on the surface. Thus, this represents the habitable region of the exoplanet. In the rest of the planet, especially in the presence of moving molten rock. In addition, I used a very simple atmospheric model on the planet’s surface with no winds or vertical layers of different temperatures and different chemical compositions.

However, despite its simplicity, the simulation reveals a variety of features. The most important is perhaps the prediction of the presence of underground oceans, like those thought to exist beneath the surface of Jupiter’s satellite Europa. The fact that Proxima Centauri b is a tidally-locked planet is essential because this creates an asymmetry in the dissipation of internal heat between the two hemispheres of the planet.

The simulation also makes it clear that for a planet in the habitable zone, such as Proxima Centauri b, it is also very important to consider its orbital parameters. In particular, on a tidally-locked planet, only the mid-latitude regions of the illuminated face can be considered habitable. However, life sustained by liquid water may also be possible on the dark side in conjunction with the presence of underground oceans.

In this simulation, the presence of underground oceans appears at about 15 Gyr. This time is larger than the current age of the Universe. However, this is still compatible with the expected lifetime of low-luminosity red dwarfs stars, like Proxima Centauri, which will take trillions of years to burn through their fuel.\textsuperscript{15}

It cannot be excluded that an ocean below the surface could form earlier if the cooling is more intense because of a lighter atmosphere or a greater distance between the planet and the star.

In the future, I would like to further develop PLANET CODE including developing in the program a physical simulation of the atmosphere and the water bodies.

\section{Acknowledgments}
I would like to thank Dr. Giambattista Aresu for following me during the project and for his valuable suggestions during the writing process. I would also like to express my gratitude to my parents for their support. Without their encouragement, this idea would not have been possible. I am very grateful to the referees for the useful and constructive comments that helped to improve this paper.
References


Author
Giulia Murgia is a senior at Asproni High School, Iglesias, Italy. She plans to pursue a major in physics. She is passionate about space exploration and astrophysics. She is amazed at how physics and mathematics combined produce predictive power that can be used to interpret the complexity of the world.
Introduction

Breast cancer is the most common type of cancer in women besides skin cancer and the second most common cause of death after lung cancer. Worldwide, there are about 1.7 million cases of breast cancer diagnosed every year, with approximately one new case detected every 18 seconds.¹ It is estimated that 281,550 women in the US will be diagnosed with invasive breast cancer and 49,240 women will be diagnosed with non-invasive breast cancer this year.² There are several factors that may raise a woman’s risk of developing breast cancer such as increasing age; personal history of breast cancer; family history of breast cancer; nongenetic, Nonmodifiable Risk Factors like race, early menarche/late menopause, etc.; and Modifiable Risk Factors like hormone use, tobacco, alcohol, and nutrition.³

Breast cancer begins when the healthy cells in the breast grow out of control and form a mass or sheet of cells called a tumor. Mutations affect the function of tumor-suppressor genes and/or oncogenes. The ‘multiple hit model’ of cancer formation hypothesizes that a single cell must have several damaged genes and receive a series of mutations that build up over time in order to become cancerous. This model shows why cancer is so hard to treat: these multiple genes mutate and interweave to create a complex and heterogeneous population of tumor cells.⁴ In a sense, one tumor can be considered multiple diseases. This is why scientists are trying to harness the immune system to fight cancer—it is the best possible defender against such a varied onslaught. Tumor suppressor genes can slow down cell division, repair damaged DNA and make cells go into apoptosis (programmed cell death). Cancer can be formed when these tumor suppressor genes do not work properly, causing the cells to grow out of control.⁵ Proto-oncogenes aid with cell growth; however, when these genes mutate/change or overproduce copies, they can become permanently activated, from (proto-oncogene to oncogene) when they are not supposed to be. This causes the cell to grow out of control, leading to cancer.⁶

In this paper the breast cancer subtype Luminal A, also known as HR+/HER2- will be the focus. It is the most common subtype, making up 68% of all breast cancers.⁷ Hormone receptor (HR) positive means that the tumor cells have receptors for the hormones estrogen or progesterone, leading to the development of HR+ tumors. Human epidermal growth factor receptor 2 negative (HER2-) means that the tumor cells do not have abnormal levels of HER2 proteins.⁸

Prognosis for luminal type A is better than other breast cancer subtypes since it has high hormone receptor expression, negative HER2 expression (does not grow fast and is not likely to spread to the lymph nodes quickly) and a low proliferation rate (slowly dividing cells).⁹ While these factors give those with this cancer a better prognosis than most, it is notable that this subtype is the most commonly diagnosed. This combination of factors makes it a good candidate for testing experimental treatments to demonstrate the efficacy of new cancer technology. Treatment for breast cancer represents a great strain on the resources of public health, so essentially eliminating it would act as a boon to the system with cascading benefits for everyone.

Current Treatments for Luminal A Breast Cancer and Their Shortcomings:

Treatment decisions for breast cancer rely on immunohistochemistry markers as well as nodal status, tumor grade, and tumor size. Depending on these factors, one or a combination of these three treatment routes will be decided upon:

ABSTRACT: Luminal A breast cancer is the most common subtype of the most common cancer in women. Current treatments using surgery and chemotherapy have greatly improved patient outcomes in recent decades but are still far from perfect. mRNA vaccines, which instruct the body to create specific proteins, hold the best promise going forward as a “cure” for luminal A breast cancer. Immunological therapies are emerging in solving problems that chemotherapy, surgery and targeted therapies cannot. The ultimate goal of immunotherapy would be to help the immune system recognize and destroy tumor cells as well as make antigens. In this paper, mRNAs potential for addressing the heterogenous nature of tumors, envisioning how it might be made to effectively elicit an immune response are explored; in particular, how it can be modified quickly to change which mutated genes it targets. This paper also discusses the limitations of this emerging technology and the difficulties of evaluating its efficacy when it has yet to enter clinical trials. Finally, this paper concludes by offering thoughts on how its development might be accelerated, predictions on effective targets, and likely dates for such a vaccine to debut as a real-world treatment.

KEYWORDS: Biomedical and Health Sciences; Cellular and Molecular Biology; Cell, Organ and Systems; Immunology; Cellular Immunology; Oncology; Breast cancer; mRNA.
Surgery:

Early-stage breast cancers are often easily treatable with surgery, but undetected and large tumors progressing to later stages quickly become more difficult to treat this way.⁹ (This is because of the difficulty inherently involved in determining whether all cancer cells of large tumors have been successfully removed) Surgery is performed to remove as much cancer as possible. A mastectomy is a type of surgery where the entire breast or both breasts (double mastectomy) are removed. In contrast, a lumpectomy/breast-conserving surgery is one in which only the specific part of the breast with cancer is removed. An advantage of this type of surgery is that a woman can keep most of her breast, though she will often need radiation therapy as well. For early-stage breast cancer, breast-conserving surgery with adjuvant radiotherapy works best. Navigating the timeline (immediately, 6 months after, etc.)⁸ of and complications around reconstructive surgery adds an additional element of difficulty with this treatment pathway.

Chemotherapy:

Chemotherapy controls the cancer by restraining its spread, making it grow more slowly and killing cancer cells that may have metastasized (When the cancer has spread from one part of the body to another, presenting further complications to treatment that must move beyond the localized). Neoadjuvant chemotherapy (before surgery) might be given to patients to shrink the tumor so that it may be removed with less extensive surgery. Usually used to treat locally advanced cancer, neoadjuvant chemo can lower the risk of recurrent cancer. Adjuvant chemotherapy (after surgery) is given to kill cancer cells that were not removed in surgery or that might have spread and cannot be seen. This also lowers the chance of recurrent cancer. Chemo is often most effective if combinations of drugs are used: anthracyclines such as Adriamycin and Ellence can be utilized alongside taxanes such as Taxol and Taxotere for both adjuvant and neoadjuvant chemo, while more aggressive anthracyclines like Doxorubicin, pegylated liposomal doxorubicin, and Epirubicin are combined with taxanes like Abraxane to treat metastasized breast cancer. Side effects of these drugs include menstrual changes and fertility issues, heart damage, nerve damage (neuropathy), hand-foot syndrome, chemo brain, and fatigue.¹⁰

When looking specifically at luminal A and how it reacts with chemo, studies fail to show benefit for patients. It still remains to be determined if chemotherapy has clinical significance, especially for patients with positive lymph nodes—meaning that the cancer has spread from the original tumor to the surrounding areas but has not metastasized yet. One review compared 5-year survival-rate differences among patients with Luminal A tumors who received and did not receive chemotherapy: the rate for patients at high clinical risk and low genomic risk without distant metastases who received chemotherapy was 1.5% higher, while the absolute difference among patients at low clinical risk and high genomic risk was only 0.8% higher.¹¹ As nearly all luminal A-like tumors fall within the low genomic risk category, these results demonstrate little benefit of adjuvant chemotherapy with this cancer subtype. Additionally, studies have shown that high-risk premenopausal patients with other breast cancer subtypes had beneficial results from receiving chemotherapy, while Luminal A patients did not experience such gains. Patients may even be harmed by the toxicities of the chemotherapy.

Targeted Therapies:

The main shortcoming of targeted therapies is that they are still very early in development. They rely on administering drugs specific to the patient’s cancer subtype, generally in combination with other treatments, to perform actions like tumor shrinkage.¹²

For decades, endocrine therapy (fulvestrant-FASLODEX, letrozole-FEMARA) has been the cornerstone for management of luminal breast cancer. Despite the substantial benefit derived by patients from endocrine therapy, primary and secondary resistance to endocrine therapy are serious clinical issues.¹³

Pathways involved in the biology of endocrine resistance have been well studied in the last two decades, leading to the development of several classes of targeted agents that have been approved. Today, in the advanced setting, three distinct classes of targeted agents are approved for use: mTOR (everolimus-AFINITOR), CDK 4/6 (palbociclib-IBRANCE, ribociclib-KISQALI, abemaciclib-VERZENIUS) and PI3K inhibitors (alpelisib-PIQRAY).¹⁴

CDK 4/6 inhibitors are the most important of these, having changed the natural history of this disease in the advanced setting and being currently under study in the early setting.¹⁵

In short, combining endocrine and targeted therapies has changed the landscape in advanced disease; in early disease, it is possible to have a similarly large impact, particularly in patients with higher risk of relapse. Moreover, experimental targeted drugs are in development such as AKT/PTEN inhibitors (Ipatasertib and capivasertib).¹⁶

Efforts to develop new agents with SERD (Selective Estrogen Receptor Degrader) properties with potent antiestrogenic activity in breast tissue have led to the discovery and characterization of second and third generation SERDs, orally bioavailable, some of which are now undergoing clinical evaluation. These include Elacestrant, SAR439859, GDC-0810 and AZD9496, among others.¹⁷

Finally, it is noteworthy that the development of targeted-agent combinations for luminal disease faces several challenges. Firstly, outcomes are already exceedingly good for most patients. Additionally, late recurrences—after five years—are a possibility for at least 20 years, which translates into a very long period of follow-up looking for events to occur. Lastly, side effects of treatment considered acceptable in advanced disease, such as such as alopecia, may not be tolerated by patients with early disease.¹⁸

The Potential of mRNA Vaccines for Treating Luminal A Breast Cancer:

While there are many potential new treatments being explored for cancer, such as CRISPR gene-editing, microbiome treatments, and cell therapy, the one that holds
the most promising for luminal A breast cancer in particular is mRNA vaccine technology. There are a few types of vaccines being explored as potential cancer treatments: immune cell–based vaccines, peptide–based vaccines, viral vector–vaccines and nucleic acid–based vaccines. (At least one vaccine, inoculating teenagers against the human papillomavirus, is currently working to reduce future rates of cervical, anal, oropharyngeal, etc. cancers that are caused or contributed to by the virus.)¹⁹ mRNA vaccines in particular are classified as nucleic acid–based and hold an especially potential to protect against infectious diseases and rapid malignant cell growth.

mRNA is, in essence, a template of instructions for how to build a protein. This technology has been gaining promise as the coronavirus pandemic has brought a lot of attention to it. The two most prominent COVID–19 vaccines in the US both utilize mRNA, which has skyrocketed interest (and funding) for additional mRNA research.²⁰ Both BioNTech and Moderna are pursuing mRNA cancer vaccines.²¹ These prove to be a little trickier than an mRNA vaccine for a virus since that case deals with clearly foreign viral matter, while cancer vaccines have to train the body to identify those portions of itself that are “foreign” malignant cells.²²

How will this work? A particular sequence of mRNA will be injected into the muscle and taken up by any cells in the surrounding area. The immune system will then hopefully recognize the proteins coded for by the mRNA as foreign and mount an immune response. The result of the immune response is that if the antigens are immunogenic enough, the T–cells (white blood cells that actively seek out the cancer cells and destroy them) will learn how to recognize these cancer cell antigens as foreign. This chain of events should trigger an immune response, wherein B cells (which produce antibodies that protect us from getting infected if, for instance, a virus enters the body) will subsequently seek out similar tumor cells.²³

mRNA–based cancer vaccines can target tumor–associated antigens, expressed in cancerous cells, for example, growth–associated factors, or choosing specific antigens that are unique to malignant cells owing to somatic mutation. These neoantigens provide tumor–specific targets for developing personalized cancer vaccines.²⁴

In short: upon administration of mRNA vaccines, tumor antigens will be expressed in a heterogenous group of immune cells in order to help antigen–presenting cells activate and elicit an immune stimulation.

To ensure the mRNA reaches its destination, naked mRNA (after in vitro transcription, it is not bound to protein and vulnerable to different enzymes that are going to recognize it) can be formulated with a delivery vehicle (vehicle–loaded mRNA).²⁵ Delivery vehicles help to improve stability, RNA uptake and translatability of mRNA vaccines. Application of exogenous RNA combined with a polymeric carrier was also shown to activate the immune system by generating a local immunostimulatory environment.²⁶

mRNA vaccines are a strong choice over other vaccines for treating cancer because they are potent, safe and easily (as well as inexpensively) modifiable. The production of mRNA is simpler and easier than protein production and purification and results in a more stabilized product.²⁷ The trick is in trying to deliver the mRNA in a way that helps maximize potency while doing that safely. Potential methods for maximizing potency include modifications made to the structure of the mRNA molecule (e.g., self–amplifying mRNAs, codon optimizations, nucleotide modifications, etc.) and different means of formulation (lipid nanoparticles, peptides, polymers, etc.).²⁷ But again, a strength of working with mRNA is how readily its sequence can be changed to add or swap targeted antigens. The quickly swappable nature of mRNA vaccines is the best way to address the challenge presented by tumors’ high degree of heterogeneity.

While mRNA vaccines are providing a powerful new avenue for treating cancer, there do remain challenges to creating an effective mRNA cancer vaccine. The products are required to be the right purity of mRNA, sequenced as well as delivered and administered properly. Given the difficulty of determining what factors are specifically driving an individual’s cancer to be out of control, it is hard to pick the right protein to encode instructions for the mRNA that will elicit the needed immune–system response. This goes alongside the general issue with cancer therapies of managing the differentiation of normal tissue from cancer cells.²⁷

That said, this work is not merely theoretical: mRNA vaccines are already being developed in the treatment of other cancers (such as skin, lung, and pancreatic cancers).²⁸ Several clinical trials with mRNA vaccines against cancer are ongoing, and nonclinical research is active in this field. Noteworthy results were released in November 2020 of an mRNA mixture being injected into mice with skin and lung cancer that successfully triggered an immune response. Cytokines produced by this response were able to shrink 85% of tumors in the mice in the span of 40 days.²⁹ These cancer vaccines, produced by the same companies behind the mRNA vaccines for COVID–19, will soon be entering clinical trials with humans in an exciting next step.

Possible Components of an mRNA Vaccine for Luminal A Breast Cancer:

In light of the above, a theoretical combination of treatments that will in effect “cure” luminal A breast cancer can be proposed. In particular, a way forward using mRNA vaccines that will effectively address the inherent heterogeneity of tumors that so often impedes traditional cancer treatments can be outlined.

From the outset, picking the right target is the biggest hurdle in making the mRNA vaccine against cancer. The first potential target would involve picking a protein present in the cancer cells but not the healthy ones—i.e., the genes that are most mutated. When looking at luminal A breast cancer, PI3KCA is the most mutated gene (about 45%), followed by MAP3K1, GATA3, TP53, CDH1, and MAP2K4.³⁰ Genes like PI3KCA contain instructions to make proteins such as the p110 alpha protein (p110–α, subunit of phosphatidylinositol 3–kinase). This gene encodes a lipid kinase involved in vital signaling pathways, fundamental for cellular functions (e.g., growth, death, and proliferation). However, when PI3KCA is mutated,
there is a change in the single amino acid produced in the p110-α, leading to the production of an altered p110-α subunit that in turn makes PI3K abnormally active. This allows PI3K to signal without regulation. The unregulated signaling results in uncontrolled proliferation of cells, causing cancer. Therefore, targeting the altered p110-α in this mRNA vaccine could successfully short-circuit this entire process.

To circumvent cancer resistance observed with chemotherapy and given the heterogeneity of tumors (not every cancer cell they are in will be equally vulnerable to the use of one targeted antigen), it would be better to use two or more mRNAs. As such, this vaccine could ideally deliver not one but several different mRNA sequences encoding for distinct proteins, ideally eliciting a robust immune response towards the full spectrum of what the cancer is expressing. However, a possible drawback here is the potential for creating a situation where there is an autoimmune response. It is not desirable to encourage the immune system to attack more than just the cancer and damage healthy cells, and yet more targets inherently increases the chances for this to happen. To avoid this scenario, it may make sense to start with a single mRNA molecule and test for whether it creates a strong enough response; if it does not (as is likely), a combination of molecules would be used or one after the other could be used to see what works best. A combination of PI3CKA, MAP3K1, and GATA3 could work well, for instance, at which point it would be possible to pare down to fewer targets to determine which target or combination of targets made it so effective.

Direct administration of complex mRNA is now considered to be a fast and feasible approach. However, there are several problems with the mRNA vaccine such as limited in vivo delivery, possibly due to enzymatic degradation, and limited intracellular delivery due to the large size of the mRNA molecule compared to other payloads. This could be overcome by screening different delivery materials and formulation methods.

Another possible avenue of interest might be to combine mRNA vaccines with immune checkpoint inhibitors. Most tumors strive to evade the immune system by expressing immunosuppressive proteins on their surface, called checkpoint proteins. Immune checkpoint inhibitors work by blocking that checkpoint. This prevents the "off" signal from being sent, allowing the immune cells to kill cancer cells. For instance, high CD70 expression may inhibit the anti-tumor immune response and is a promoter in tumor progression. CD70 expression (how much protein there is) could allow the tumor cells not to trigger an immune response. By inhibiting immune response, the CD70 could work against the mRNA vaccine. As such, any ways to diminish that possibility (such as combining with different treatment options) should make this vaccine more effective.

**Discussion**

Only studies can truly determine the efficacy of new treatments, which limits the possibilities of what a literature review in this area can accomplish. And as few mRNA vaccines for cancer treatment are yet in clinical trials, they are not widely reported on in the literature—making research into this area difficult. That said, the success of mRNA vaccines in fighting the spread of the novel coronavirus signals a clear turning point in this technology. The possibility of a groundbreaking treatment here should not be denied.

With the right antigen selection and fine-tuning of mRNA delivery, payload size, etc., it is possible to unlock the full potential of the best ally each human body always had in attacking cancer cells: the human immune system. If the immune system can be properly instructed to treat cancer cells as the “foreign” bodies that they essentially are, their eradication of luminal A breast cancer tissue could end up being far more complete than what is possible with surgery and chemotherapy alone—and in a way that improves patients’ physical and psychological outcomes as treatment becomes less invasive.

This ‘hacking’ of the immune system is a new and exciting field of science that is moving at a rapid clip. With how quickly mRNA vaccines can be developed, deployed, redeveloped, and redeployed, progress should be expected quickly. The promise of mRNA vaccines (and possibly oRNA and miRNA as well) as proposed above for luminal A breast cancer is not a far-off or far-flung dream, but a promise that will likely be realized within the next decade or two.

**Conclusion**

mRNA vaccines are a highly promising route for treating or even “curing” luminal A breast cancer. They should continue to be explored at an aggressive pace, with increased funding allotted to speed them into clinical trials in the wake of their massive success in combating COVID-19.

It is likely that, in order to effectively address the heterogeneity of tumors, these vaccines will have to target multiple antigens simultaneously. This paper predicts that PI3CKA and MAP3K1 will prove the most worthwhile genes to target, and that doing so with effective payload delivery and minimal autoimmune response triggered will be possible in real cancer therapies by 2035. Particular attention should be paid to how best to combine these mRNA vaccines with existing treatments like surgery and chemotherapy as well as other cutting-edge treatments like targeted therapies.

These advances in treatments should function as a “cure” for luminal A breast cancer by 2040, with the relieved burden on the health care system as well as on women everywhere ushering in a new age of wellness and positive life outcomes. Even more excitingly, given the flexibility with which researchers can rework mRNA in the lab, any advances in mRNA vaccines for cancer should prove a ready boon to cancer researchers of all stripes—and assuredly to those looking into other perennial diseases likewise.

**Acknowledgements**

Thank you for the guidance of Daniel Wilson, my mentor from Carnegie Mellon University, in the development of this research paper.

**References**


Targeted Therapy https://www.nationalbreastcancer.org/breast-cancer-targeted-therapy.


Cristina Hernando, Belén Ortega-Morillo, Marta Tapia, Santiago Moragón, Maria Teresa Martínez, Pilar Eroles, Iris Garri do-Cano, Anna Adam-Artigues, Ana Lluch, Begona Bermejo, and Juan Miguel Cejalvo.


Cristina Hernando, Belén Ortega-Morillo, Marta Tapia, Santiago Moragón, Maria Teresa Martínez, Pilar Eroles, Iris Garri do-Cano, Anna Adam-Artigues, Ana Lluch, Begona Bermejo, and Juan Miguel Cejalvo.


- Authors
Alexandra Giuliani is a junior at the International School of Monaco. She enjoys studying molecular biology and looks forward to researching biotechnology in college.
Analyzing Consumer Behaviors and Attitudes Towards Plastic Bag Consumption in Hanoi Wet Markets

Jinmoo Yoo
Hanoi International School, 48 Lieu Giai, Hanoi, Hanoi, 100000, Vietnam; jinmoo.yoo@gmail.com

ABSTRACT: This research investigates consumers’ stated attitudes and behaviors towards plastic bag consumption in Hanoi wet markets. The project leveraged a stated preference survey (SPS) and background demographic survey of the consumers at two wet markets to run a linear regression and determine the “value” that the general Vietnamese population places on plastic bags compared to alternatives. Altogether, the demographic survey gathered 163 responses and SPS 235 responses. The SPS provided survey respondents with four different options to store the vegetables they bought in the wet market: plastic bag, paper bag, eco-friendly tote bag, or a reusable bag. Four different attributes (price, volume item can hold, weight item can hold, and the number of times item can be used) and their respective levels were presented. The analysis (p<0.05) showed that price and functionality are the two most important actors in deciding whether a consumer would use a plastic bag or an alternative at the wet market. For each additional unit kilogram that an item, could hold, the probability of a person purchasing that item decreased by 3.21%. On the other hand, a unit increase in price (measured in VND) decreased the probability that a person would purchase the item by 0.0001178%. Ultimately, despite the current excessive use of plastic bags, the general Vietnamese demographic (91%) demonstrated a shift or a desire for a shift towards the adoption of plastic alternatives.

KEYWORDS: Earth and Environmental Sciences; Environmental Effects on Ecosystems; Data Science; Behavioral Economics; Plastic Alternatives; Linear Regression; Effective Preferences.

Introduction

Plastics are crucial to modern society due to their versatility, affordability, and durability. As a result, the production of plastics has increased exponentially within the past 70 years: while only 2 million tonnes of plastic were produced in 1950, 350 million tonnes of plastic were produced in 2020.¹ Further research shows that plastics contribute to climate change² and can be detrimental to human health for multiple reasons.³

Plastic pollution itself threatens human health because humans can consume microplastics when eating seafood. Though over 90% of the microplastics humans ingest are digested properly, they have the potential to enhance respiratory responses and disrupt the gut microbiome due to the toxicity and chemical transfer of pollutants from plastic particles.⁴

Hanoi, the capital of Vietnam, is home to 8 million people with a GDP growth rate of 7.62% as of 2019.⁵,⁶ Unfortunately, but in line with its urbanization rates, booming economy, and limited government enforcement, Vietnam ranks as the 7th largest contributor to plastics pollution globally while ranking 15th by population. Likewise, Vietnam’s consumption of plastics continues to grow as it develops, generating over 1.8 tonnes of plastics per year.⁷

There have been multiple government efforts in order to reduce the excessive plastic consumption. For example, on April 11th, 2013, the government proposed a plan to improve environmental pollution control regarding the use of non-biodegradable plastic bags. The policy set a limit on the total number of plastic bags used and a newly devised waste management system that would take place by 2020.⁸ Moreover, on May 7th, 2018, the government adjusted its initial national strategy to improve solid waste management: by 2050, Vietnam will rely 100% on eco-friendly plastic bags in shopping malls and supermarkets as an alternative to non-biodegradable counterparts. Furthermore, the government has stated it will also restrict the production and imports of plastic bags starting in 2026.⁸

Vietnam has an increasing middle class and adopts a consumerist lifestyle as a means of enhancing one’s identity and social status.⁹ Therefore, when an updated version of a product is introduced, people will quickly adapt and switch to using the updated product. This study operates under the belief that consumerism can be leveraged to alter plastic consumption in Vietnam: an introduction of new plastic alternatives for free or at low-cost could drastically reduce the number of plastic bags used, as it is easy to make customers adopt the alternative option.¹⁰

Pre-existing data on Vietnam’s plastics consumption, particularly within wet markets, is elusive, but there is still some information about wet markets in general that is important. In Vietnam, wet markets are a traditional open-air, food market that sells animals, meat, fresh produce, and various other products. Furthermore, despite online and drop shipping becoming more common especially during the COVID-19 outbreak, wet market sales continue to increase rapidly, growing by 4% in 2020 alone, reaching a total revenue of VND 1,027 trillion [$44.90 million].¹¹,¹²

According to initial research, the three main reasons why people still use wet markets are due to the freshness of the products, interpersonal relationships, and the context of the situation.¹² Wet markets are a pillar of Hanoi consumerism because the middle class perceives that they sell cheaper...
fundamental reasons behind plastic bag use by examining the attitudes of consumers towards plastic bags and their alternatives. This is important as we are able to present a solution that directly resolves the concern that the consumers have expressed in regard to using plastic alternatives. In addition, it shows the possibility of the consumers adopting alternatives.

Methods

We sampled two mostly homogeneous wet markets within the same district bearing similar socioeconomic attributes: Linh Lang Market and Ngoc Ha Market. Located in Ba Dinh district, both markets sell various items, from common wet market products such as produce, tofu, freshly butchered pork, live fish, roast duck, flowers, and a variety of “to-go” juices.

By choosing markets that are located in the same district and only a 6-minute drive apart, our team attempted to minimize heterogeneity and other variables among the demographics of those visiting the markets, and thus the goods wanted within the same middle to wealthy income socioeconomic level. This is because policy makers, press, and social media influencers and market drivers reflect consumer trends among everyday trends and potential priming to ideas related to environmental sustainability and plastic alternatives.

Methodology of Demographic Survey (n=163):

Our team of four Vietnamese-speaking students conducted a 12-question-survey on July 3rd, 2021, and July 11th, 2021 from 6:30 A.M. to 9:00 A.M. This survey investigated the predominant reasons why people visit wet markets, why people use plastic or recyclable bags, what people do with plastic bags, and – ultimately – if people are willing to start using a plastic alternative in Hanoi, Vietnam. Figures 1 and 2 below show the questions asked to the wet market visitors in Vietnamese and are translated below in English.

Figure 1: Background survey first page.

We conducted the surveys at the streets bisecting the markets and in middle of the markets, and surveyors read each question and possible answer choices one-by-one to each respondent, resulting in high response recording fidelity. Due to the sudden surge of COVID cases in Hanoi on July 8th, we took precautionary measures by approaching the consumers and anyone with masks on at all times.

DOI: 10.36838/v4i3.6

Figure 2: Background survey second page.

Methodology of Stated Preference Survey (n=235):

Due to the lockdown policy implemented by the Vietnamese government on the 24th of July, 2021, our team did not collect what we felt statistically relevant for an A/B test at the time. Therefore, we ran a stated preference survey on Hanoi consumers who visit wet markets either regularly or once a month.

We presented the survey respondents with a scenario, four options that they could choose, and their respective attributes and levels, which are all shown in the paragraph and below (Figure 3).

“You are a customer shopping at a local wet market. The vendor is a friendly 40-year-old man who sells vegetables (lettuce, cucumbers, cilantro, eggplants, etc.). You have the option to take a plastic bag, paper bag, eco-friendly cloth bag or a reusable container. The descriptions for each bag are shown in the image below. Assuming all the following characteristics are true, which option would you pick?”

Figure 3: Stated preference survey options, attributes, and levels.

To begin with, the situation presented included “friendly 40-year-old man who sells vegetables” as he is an actual vendor at the wet market that our team has interacted with before. In addition, the scenario presented the respondents with the same products that the vendor sold: “lettuce, cucumbers, cilantro, eggplants, etc.” Lastly, it was made clear that the respondents should assume that all the following characteristics are true. This made sure that all consumers had a starting point in the decision-making process, allowing...
the team to draw more accurate conclusions. The survey accepted responses between August 1st, 2021, and September 1st, 2021.

The current study utilized discrete, binary numbers to represent the respondents’ actions (0 representing that they declined a particular option and 1 representing that they chose a particular option). This value was set as the left-hand value (LHS Value). Using the LHS Value allowed our team to record the consumers’ behaviors thoroughly: each respondent had four rows of data and only one reflected the choice they had made. The information was compiled into an Excel database. Table 1 shows how the Excel data is formatted.

### Table 1: Excel database format.

<table>
<thead>
<tr>
<th>ID</th>
<th>LHS Value</th>
<th>Price</th>
<th>Volume</th>
<th>Weight</th>
<th>Times</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>30000</td>
<td>12.5</td>
<td>20</td>
<td>150.0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>2000</td>
<td>3.0</td>
<td>5</td>
<td>5.0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1.5</td>
<td>3</td>
<td>2.5</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>100000</td>
<td>2.0</td>
<td>5</td>
<td>150.0</td>
</tr>
</tbody>
</table>

With the created database, an OLS regression test was run to examine the different correlations between the consumers’ decisions and other factors or variables such as price and weight the items could hold. The dependent variable was set as the LHS Value.

Ultimately, the team hypothesizes that approximately 85% of the consumers surveyed at the wet market would use plastic bags as their main method of storing products. Additionally, it was predicted that the two main factors affecting consumer decision on plastic bag alternatives would be price and availability. With this study, the team hopes to recommend a policy that effectively promotes sustainable habits and decisions for the local consumers.

### Results and Discussion

#### Respondents’ Demographics:

Table 2 visualizes 136 respondents of the initial survey’s demographics, which includes their age group, assigned (biological) sex, their frequency in visiting the wet market (times per week), and whether they live in the same district as that of the wet market. Respondents included 97 women and 39 men. Ages ranged drastically from children (0-14 years old) to the elderly (65+ years old) where the most common age group was 25 to 40 years of age, accounting for 31.62% of the respondents. 88.18% of the respondents lived in the same district as that of the wet market they visited.

Figure 4 shows what people tend to purchase in the wet markets. It can be observed that the female respondents outnumber the male respondents by 2.5:1. Moreover, there was a total of 130 purchases for vegetables, 121 purchases for meats and poultry, 88 purchases for fruits, 73 purchases for seafood, 64 purchases for tofu or other wet products, 43 purchases for dried goods, 26 purchases for flowers, 14 purchases for sundries and household goods and 2 other purchases.

### Table 2: Summary of demographic information.

<table>
<thead>
<tr>
<th>Category Group</th>
<th>Categories</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>Female</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>39</td>
</tr>
<tr>
<td>Age Group</td>
<td>0-14</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>15-24</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>25-40</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>41-55</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td>56-65</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>65+</td>
<td>21</td>
</tr>
<tr>
<td>Residence in the Same District</td>
<td>Yes</td>
<td>121</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>15</td>
</tr>
<tr>
<td>Market Visit Frequency (Times per Week)</td>
<td>1</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>7+</td>
<td>4</td>
</tr>
</tbody>
</table>

### Figure 4: Items purchased at the wet market.

**Plastic Consumption by Wet Market Customers:**

We asked people why they use plastic bags, what they do with the plastic bags after use, and whether they were willing to adopt a plastic-alternative.

Firstly, as represented by Figure 5, 55.49% of the total respondents stated that they use plastic bags out of convenience, while 17.65% of respondents stated that they use plastic bags out of their habitual practices. Only a mere 4.12% of the respondents do not use plastic bags at all. This low number
demonstrates how only a low percentage of the consumers in Hanoi wet markets are currently using recyclable bags.

43.48% of survey respondents predominantly use their leftover plastic bags to store other trash while 35.87% claim to simply dispose of the plastic bags directly into the waste stream. The third most common use for leftover plastic bags is as storage for other items (15.85%). Only a mere 1.64% stated that they do not use plastic bags and 0.54% of respondents reported reselling plastic bags to plastic recyclers.

Consumers’ Willingness to Adopt Non-Plastic Alternatives:

Responses to the question of whether respondents were willing to adopt non-plastic alternatives or not are illustrated by Figure 6. The three available answer choices for this question were “yes,” “yes, provided my concerns do not apply” and “no.” The possible concerns were those mentioned in the question that asks consumers on why they use plastic bags: there is a lack of alternatives, the alternatives are too expensive, etc.

As shown in Figure 6, approximately 50.74% of the respondents have said they are willing to use a non-plastic alternative. In addition, 44.12% of the respondents stated that they will likely use a non-plastic alternative given that their concerns do not apply. Lastly, only a mere 5.15% of the customers stated that they are not willing to use a non-plastic alternative.

Despite the current excessive use of plastic bags in Hanoi wet markets, the local consumers have demonstrated a high interest in the adoption of non-plastic alternatives. Hence, the consumers are at an ideal situation where they will likely adopt a non-plastic alternative when they become easily accessible, cheap, or satisfy their concerns.

Analysis of Stated Preference Survey:

Results indicated that 63% of the respondents chose the tote bag, 22% chose the paper bag, 9% chose the plastic bag and 6% chose the recyclable container. An ANOVA regression was conducted to validate the accuracy of the study. The test showed that the database had a final F significance value of 1.32*10⁸¹. First, the team ran OLS Regression on the LHS value against other variables, such as price, volume, weight, and times (Table 3).

<table>
<thead>
<tr>
<th>Categories</th>
<th>Values</th>
<th>Coefficient Estimate</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price</td>
<td>Intercept 0.2889</td>
<td>9.4856E-51</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Slope -1.178E-6</td>
<td>0.0007221</td>
<td></td>
</tr>
<tr>
<td>Volume</td>
<td>Intercept 0.01423</td>
<td>0.4188</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Slope 0.04964</td>
<td>2.882E-65</td>
<td></td>
</tr>
<tr>
<td>Weight</td>
<td>Intercept -0.01548</td>
<td>0.4177</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Slope 0.03218</td>
<td>9.324E-63</td>
<td></td>
</tr>
<tr>
<td>Times</td>
<td>Intercept 0.02001</td>
<td>1.565E-13</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Slope 0.001302</td>
<td>9.406E-12</td>
<td></td>
</tr>
</tbody>
</table>

Then, RStudio was used to create linear models and plot correlations between the consumers’ choices and other variables. First, the LHS value was graphed against the price of the item. According to the linear model (p = 0.000722), the probability of someone choosing a free item is 28.89%. However, for each unit increase in price (measured in VND), the probability of someone choosing that item decreases by 0.0001178%. Therefore, the probability of someone buying an item that costs 100,000.00 VND is reduced to 17.11%. This is illustrated by Figure 7.

Furthermore, the team also investigated whether or not there exists a correlation between the consumers’ choices and the weight/carrying capacity of an alternative use.

The following linear regression showed functionality as another key factor potentially influencing consumers’ decisions: for each additional unit kilogram that the item was able to hold, the probability of a purchase of that item grew by 3.21% (p = 9.3235E-63). The findings are illustrated by Figure 8.
The research discussed in this article aimed to better understand attitudinal preference in wet markets and other mini grocers in Hanoi’s urban core for changing consumer behavior. We hope to utilize this research for further investigation as well as to help spread awareness in Vietnam to some of the problems, solutions and general attitudes surrounding adoption of pro-sustainability behaviors that function and have merit. Finally, eliminating single-use plastic bags is a topic worth considering in its own. The consequences of not changing behaviors in the world’s largest markets means vast populations will continue to suffer from a more toxic local environment and global ecosystem—much of it carried out by our own hands. This would be tragically ironic since simple, cheap solutions already exist.

## Acknowledgements

I would like to thank my advisor Andrea Dunchus of Duke University for constant guidance and support and my fellow investigators at the Hanoi Urban Lab: Hoai Anh Le, Hai Nguyen, Truong Hoang Ha, Thuy Linh, and Lam Dinh for their tireless effort without whom I would have been able to accomplish little of this.

## References


## Conclusion

The demographic and stated preference survey revealed that the general Hanoi population predominantly uses plastic bags. However, results indicated that Vietnamese citizens are also open to transitioning to eco-friendly alternatives through a 91% affirmative response toward plastic alternatives. According to the on-site background survey, a main concern in adopting a plastic-alternative seems to be in how consumers are habitual in their plastic bag usage. However, the preference survey showed that 90.63% of the consumers exhibited a tendency to prefer non-plastic alternatives such as tote bags due to their robust and reusable nature. Hence, consumers have a tendency to prefer non-plastic alternatives such as tote bags due to their robust and reusable nature. This insight is useful for considering policy and marketing levers to increase alternative product use.

Eco-friendly alternatives like paper bags, tote bags, or reusable containers are responded to encouragingly as compared with a plastic bag when shopping at a wet market. However, this stated affinity in the preference survey does not reflect an actual willingness and follow through to change a behavior pattern. In reality, single plastic bag use is still the overwhelming (98.36%) norm.

All in all, considering that 91% of the respondents chose a non-plastic alternative in the stated preference survey, one potential solution the government could implement is a subsidy for the eco-friendly bag producers. This would increase the quantity of recyclable bags supplied to Vietnam’s domestic consumers and decrease the unit price of recyclable bags, which effectively targets the issue of price and availability.

To determine the true preference of respondents or to indicate an intent in their actual behavior, the next step is to complete an A/B testing method at wet markets and combine these with Google trends or a third observational study. This includes prototyping designs including setting up a stall that sells non-plastic alternatives for a period of time. This would allow consumers to potentially gain insights about such eco-friendly products; gain first-hand experience shopping with and using these alternatives at no risk, and offer corrective or confirmative bi-dimensional data to the results of our two surveys.

### Authors

Jinmoo Yoo is a senior at Hanoi International School. He aspires to leverage data science to devise private market solutions and smart government policies on pressing environmental, social, and business issues. Jinmoo hopes to major in data science or operations research with a minor in environmental sciences/engineering.
Identification of fimS Regulator Motifs in K. pneumoniae and E. coli to Advance Antivirulence Treatment of Type 1 Pili

Sarah Joseph

Parkway Central High School, 369 N Woods Mill Rd, Chesterfield, MO, 63017, USA; joseph.sarah2022@gmail.com

ABSTRACT: UroPathogenic Escherichia coli (UPEC) and Klebsiella pneumoniae are the two main causative agents of urinary tract infections. Furthermore, these uropathogens are rapidly developing multi-drug resistance. However, as type 1 pili (filament-like proteins) are critical virulence factors within both species, antivirulence treatment targeting the pili-encoding fim operon is a potential alternative to antibiotics for treatment. We used MEGA XI Mass Alignment Sequencing and phylogenetic analysis to identify conserved nucleotide regions among strains within select clades of UPEC and K. pneumoniae. From a sample of fim operon regulators, we identified statistically significant protein-binding motifs among the conserved regions using the Multiple Em for Motif Elicitation (MEME) Tool. We found the binding domains for the following regulator proteins to be highly conserved in the fimS region of E. coli reference strains: CitT, CovR, GlnR, IHF, RotA, XylR, SigW, PucR, Fur, MetR, and OxyR. CTSR and Hpr were completely conserved in K. pneumoniae reference strains, and PhoP is highly conserved in both species. These proteins are potentially associated with the pathogenicity of UPEC and K. pneumoniae. Together, these results indicate that antivirulence treatment targeting these regulators can be advanced. Future studies could conduct an electrophoretic mobility shift assay for confirmation of regulators’ binding regions and genetic analysis through motif deletion/mutation to determine the regulators’ functions.

KEYWORDS: Computational Biology and Bioinformatics; Genomics; Escherichia coli; Klebsiella pneumoniae; Type 1 pili; fimS.

Introduction

Urinary tract infections (UTIs) are common, recurrent infections primarily caused by UroPathogenic Escherichia coli (UPEC) and Klebsiella pneumoniae, which are Gram-negative, facultative anaerobic, rod-shaped bacteria. UTIs remain the most common outpatient infections with a lifetime incidence of 50-60% in adult women. They are classified on the basis of infection site: infection of the bladder/lower urinary tract is termed cystitis, infection of the ureter is termed urethritis, and infection of the kidney is termed pyelonephritis. If left untreated, the spread of uropathogens from the kidney into the bloodstream may lead to septicemia.

K. pneumoniae and UPEC strains are distinct from other pathotypes of E. coli due to their production of several virulence factors that enable pathogenesis specifically in the urinary tract. To elaborate, both bacteria form biofilm-like masses called intracellular bacterial communities (IBC) and disperse into epithelial cells to initiate further rounds of UTIs. Other notable virulence factors are type 1 pili, whose structure and corresponding operon, termed “fim operon”, are highly similar to that of K. pneumoniae (Figure 1). Type 1 pili are filament-like proteins that extend from the membranes of E. coli and K. pneumoniae and are vital for the invasion of the urinary tract. The pili terminate in the adhesin protein FimH, which adheres to O-mannosylated uroplakins (glycan-decorated transmembrane proteins embedded in the urothelium) and allows bacteria access to the urothelial niche, thus propagating infection. FimH-mediated adhesion to uroplakins is critical to UPEC and K. pneumoniae establishment and propagation of a UTI.

In both UPEC and K. pneumoniae, the expression of the fim operon is phase variable, mediated by an invertible promoter region termed fimS. It is located upstream of the sequence of structural genes from fimA to fimH, encoding the main structural subunit (FimA) and adhesin terminator protein (FimH), respectively (Figure 1). Periodic inversion of fimS into the phase ON orientation occurs so that transcription may occur from 5’ to 3’, which is mediated by recombinases FimB and FimE. Specific to the K. pneumoniae fim operon, fimK is a regulator gene located downstream of fimH. It inhibits type 1 pili production by reducing the transcription of the operon, requisite during pathogenesis of pneumonia.

Suppression of type 1 pili ensures that the bacteria are not detected by the immune system.

In conformity with fimbriae-dependent uropathogenesis, extensive study regarding the regulation of the fim operon has been conducted. Previous research concerning the inhibitory properties of urine on the function of the fim operon in E. coli shows that the growth of E. coli in urine blocks FimH function and induces the fimS phase OFF orientation. However, once FimH has adhered to human bladder cells, it reverses the inhibitory properties of urine on production type 1 pili so that fim expression turns on and remains ON, regardless of its growth in urine. This reversal highlights antivirulence treatment targeting pili specifically as particularly effective against UTIs because it is more effective than the temporarily inhibitory properties of urine. Thus, the present study aims to find putative regulator proteins of the fimS region in order to provide a basis for future research concerning the proteins’ relation to the pathogenicity of E. coli and K. pneumoniae.
well as treatment that targets protein regulation of the region itself.

With a pangenome of approximately 15,000 genes, UPEC are genetically heterogeneous; *K. pneumoniae* are also differentiated, but to a smaller extent. Given the genetic diversity across UPEC and *K. pneumoniae* strains, the presence of motifs (conserved regulator protein-binding sequences) in the *fimS* region, unique to different clades of UPEC and *K. pneumoniae*, can provide insight into the genetic regulation of the *fim* operon. In the present study, we aim to characterize the two *fimS* regions using various reference strains of published *E. coli* and *K. pneumoniae* genomes. Specifically, we aim to utilize phylogenetic and motif analysis to determine how the *fimS* region of all *E. coli* compares to that of *K. pneumoniae* and how it compares between clades of UPEC and within each species itself.

**Methods**

**Nucleotide BLAST of *E. coli* UTI89 *fimS* region:**

The Nucleotide Basic Local Alignment Search Tool, or Nucleotide BLAST, is a bioinformatics algorithm that compares and finds statistically significant similarities between a query DNA sequence and a database of genomic sequences.⁷ We sourced reference strains of UPEC from each of the following five clades of *E. coli*: B2, D, E A, and B1 (Table S1).⁸,⁹ Given the relatively lower genetic diversity of uropathogenic *K. pneumoniae* strains, the analyzed *K. pneumoniae* strains were all classified as the members of the same clade; thus, we did not conduct phylogenetic analysis of the *K. pneumoniae* strains in this study. For both pathogens, we defined the *fimS* sequence as the intergenic region between *fimE* and *fimA*.

To conduct a phylogenetic analysis of *fimS* regions among each UPEC clade and strains of *K. pneumoniae*, we first used the Nucleotide Database in the National Center for Biotechnology Information (NCBI) Search Tool to determine the accession number for each reference strain.¹⁰ Into the Nucleotide BLAST algorithm, we then entered the known FASTA sequence of the *fimS* region in *E. coli* UTI89 as the query sequence and the accession number for each reference strain as the subject sequence for comparison. The same was done with the *K. pneumoniae* strain C3091. For each blast, the *fimS* sequences were entered as either the ON or OFF orientation, dependent on the orientation of the subject sequence. We recorded the resulting FASTA sequences of the entire intergenic region from the end of *fimE* to the start of *fimA*, containing the *fimS* region, for each reference UPEC and *K. pneumoniae* strain.

**Alignment of *fimS* Sequences Per Clade and Phylogenetic Analysis:**

Upon recording the *fimS* FASTA sequences for all pathogen variants, we used Multiple Sequence Alignment (MSA) tool ClustalW, embedded in the Molecular Evolutionary Genetics Analysis 10 algorithm (MEGA 11), to automatically align the *fimS* sequences for each clade of *E. coli* and all *K. pneumoniae* strains under default parameters. We then used MEGA 11 to construct a phylogenetic tree inferred by the Maximum Likelihood method and the Tamura-Nei model. The phylogenetic trees with the highest log likelihoods were constructed, indicating the highest level of fit of the tree to the sample of alignment sequence data. Initial trees for the search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Tamura-Nei model, and then selecting the topology with superior log likelihood value.¹¹,¹²

On the basis of evolutionary distance, we formed phylogenetic cluster groupings within each clade’s tree and the *K. pneumoniae* tree (Figure 2). We then selected a randomized representative consensus sequence from each grouping, which we used to identify the recurring regulator-protein binding sequences (motifs) found in the sequence. Cluster analysis of pairwise evolutionary distances between all UPEC and *K. pneumoniae* isolates revealed two to four clusters within each phylogenetic tree. Outgroups in clades A, B1, and B2 were each assigned to their own cluster (Figure 2B, C, D). We then identified one reference strain from each cluster as its evolutionary representation (for use in motif identification).
Figure 2: Phylogeny Trees of Studied K. pneumoniae strains and UPEC Clades. Trees shown: K. pneumoniae strains (A), UPEC Clade A (B), UPEC Clade B1 (C), UPEC Clade B2 (D), UPEC Clade B2 (E). The trees were inferred by using the Maximum Likelihood method and Tamura-Nei model. The trees with the highest log likelihoods are shown. Initial trees for the search were obtained by applying NeighborJoin and BioNJ algorithms to a matrix of pairwise distances estimated using the Tamura-Nei model, and then selecting the topology with superior log likelihood value. The trees are drawn to scale, with branch lengths measured in the number of substitutions per site. On the basis of evolutionary distance, phylogenetic groupings (clusters) were formed for each tree, boxed in green. Representative strains of the cluster genome, selected at random for motif identification, are highlighted in yellow.

Motif Comparison and Genotype–virulence Association:

We used the Motif Alignment and Search Tool (MAST) embedded in the Motif Elicitation (MEME) version 5.1.1 suite to identify motifs within the consensus fimS sequences of each representative UPEC and K. pneumoniae strain. We ran the MAST algorithm for each representative strain against the PRODORIC database, constructed from a set of aligned transcription factor DNA-binding sites. We set the E-value sequence display threshold at less than or equal to 20. Upon completion, we analyzed results in the HTML format; the resulting block diagram shows the best non-overlapping tiling of motif matches on the sequence. We paired the motif matches with their associated regulator proteins and recorded the proteins/corresponding representative strains in a chart for further analysis. The position p-value of each motif was below the 0.0001 significance threshold, set by default.

Results and Discussion

Phylogenetic relatedness and cluster analysis of UPEC and K. pneumoniae strains:

The intergenic region between fimE and fimD of 55 E. coli and 45 K. pneumoniae reference strains (Table S1) were determined, aligned, and combined into phylogenetic trees inferred by the MEGA XI algorithm in order to evaluate aspects of the strains’ relatedness and genetic content (Figure 2B, C, D, E). To elaborate, four phylogenetic trees were constructed from samples of UPEC reference strains belonging to the following clades: A, B1, B2, and D. One phylogenetic tree was created for all K. pneumoniae reference strains studied due to the relatively high level of relatedness among these strains (Figure 2F). Each tree was inferred by the Maximum Likelihood method in the MEGA XI algorithm; Neighbor–Join and BioNJ algorithms were applied to estimates created by the Tamura-Nei model, and the tree with the highest log likelihood was selected. The evolutionary distance scale is based on the number of nucleotide substitutions per site, which varies among each tree.

Phylogenetic cluster analysis of pairwise evolutionary distances between all UPEC and K. pneumoniae strains revealed two to four clusters within each phylogenetic tree. Strains of every cluster, with the exception of K. pneumoniae strain 4928STDY7071137 (in the fourth cluster, see Figure 2A), maintained an evolutionary distance below 0.10 nucleotide substitutions per site.

Putative fimS regulator motifs conserved among UPEC and/or K. pneumoniae:

We identified the following putative fimS regulator target sequences to be highly conserved among the E. coli reference strains: CitT, CovR, GlnR, IHF, RofA, XylR, SigW, PucR, Fur, MetR, OxyR. Furthermore, only IHF, SigW, PucR, and PhoP binding motifs were present among all UPEC representative strains. In the K. pneumoniae representative strains, CTsR and Hpr motifs were present in all strains.

Figure 3: Motif conservation among E. coli and K. pneumoniae representative strains. Regulators associated with motifs identified by the MAST tool are plotted against corresponding reference strains (which presented a match), highlighted in gray. The grid is organized by high conservation among UPEC strains, K. pneumoniae strains, and both species; these three regions are outlined in red. UPEC strains are color-coded by clade.

Concurrent presence/absence of regulator pairs:

In UPEC strains 11128, 95–3322, and Sakai, binding motifs for regulators XylR and OxyR are concurrently absent but are present in all other strains (with the exception of only XylR being present in UPEC strain 317, an outgroup of clade B2). Similarly, the Fur and MetR motifs are concurrently absent among strains, also with the exception of strain 317, in which only the MetR target sequence is present (Figure 3). The concurrent presence and absence of the two pairs of motifs may suggest a mechanism that involves both motifs.

Discussion

Type 1 pili (filament-like proteins) are critical virulence factors within both E. coli and K. pneumoniae. Thus, antivirulence treatment targeting the pili–encoding fim operon is critical. In order to develop a putative association between pathogenicity and regulatory proteins of the fimS region, we characterized the fimS regions in E. coli and K. pneumoniae and identified regulator motifs within each species. Upon the assumption that certain regulatory proteins are conserved among K. pneumoniae, E. coli, or both, we conducted a phylogenetic analysis of sample strains and used the motif elicitation algorithm MAST, paired with the PRODORIC motif database, to identify putative protein binding motifs in the fimS region.

Upon phylogenetic cluster analysis, strains of every cluster, with the exception of K. pneumoniae strain 4928STDY7071137
4928STDY7071137 (Figure 2A), maintained an evolutionary distance below 0.10 nucleotide substitutions per site. This indicates a similar evolutionary relationship among the strains within each cluster, suggesting the fimS region is highly conserved across different strains of *E. coli* and *K. pneumoniae*. Furthermore, the formation of phylogenetic clusters allows us to identify representative strains for which we can take representative data and look at conservation of the proteins within clusters and/or highly related strains.

We were able to identify, based on Multiple Em for Motif Elicitation (MEME) results, the binding motifs for the following putative fimS regulators in the *E. coli* reference strains: CitT, CovR, GlnR, IHF, RofA, XylR, SigW, PucR, Fur, MetR, OxyR. In the *K. pneumoniae* representative strains, CTSR and Hpr motifs were present in all strains. Among both species, PhoP was present in all strains, with the exception of *K. pneumoniae* strain INF291-sc-2280225. Based on prior published works, groups have confirmed IHF (integration host factor) as a regulator of the fimS switch; specifically, binding of the IHF protein to a site immediately adjacent to fimS is required for phase-on orientational bias.⁴⁻¹⁶ We have similarly found IHF to be completely present among every *E. coli* reference strain, and it is absent among all *K. pneumoniae* strains. Based on previous studies, *K. pneumoniae* requires the inhibition of type 1 pili production during pathogenesis of pneumonia to ensure that the bacteria are not detected by the immune system.³ As IHF is confirmed to be an activator of the fimS region, we hypothesize that our *K. pneumoniae* reference strains lack this protein due to its nature as a pneumonia-causing pathogen.

Furthermore, previous studies have observed increased expression levels of the phoP gene in colistin (antibiotic) resistant *K. pneumoniae* strains, thus identifying PhoP as responsible for cell membrane modifications related with colistin resistance.¹⁷ Among avian pathogenic *E. coli* (AVEC), the deletion of phoP in *K. pneumoniae* has also been shown to reduce biofilm formation.¹⁸ Further a genetic study involving the deletion of associated motifs can confirm whether the role of PhoP in the fimS region of *E. coli* and *K. pneumoniae*, or the intergenic region between fimE and fimA, presents any relation to its roles as an activator of colistin resistance and biofilm formation.

The presence of CTSR and Hpr in all *K. pneumoniae* reference strains and none of the UPEC strains, upon further study, suggests their functions as repressors of the fimS region, which is requisite for a pneumonia-causing pathogen.⁵ Additionally, the concurrent presence/absence of certain proteins in UPEC suggest that these proteins may work together and constitute a single mechanism.

Despite the variety of diseases caused by each *E. coli* clade, the regulator motifs we identified are highly conserved among all clades (Figure 3). To elaborate, virulent extra-intestinal *E. coli* strains belong mainly to clades B2 and D, whereas the largely commensal strains belong to clade A.¹⁹ However, given that putative motifs are largely conserved among all the *E. coli* clades studied, differences in motifs present are more apparent among comparison between *E. coli* and *K. pneumoniae*.

As not all the proteins found are confirmed to be regulators, we cannot definitively associate every protein with pathogenicity. Other researchers could conduct an Electrophoretic Mobility Shift Assay with the potential binding proteins in *K. pneumoniae* and/or *E. coli* in order to confirm their binding to the intergenic region between fimE and fimA.²⁰ Furthermore, the deletion and/or mutation of potential binding sites in uropathogenic strains can be analyzed on the basis of what functions they alter and how they affect type 1 pili production. The association of a given protein with pathogenicity can then be confirmed.

By investigating regulators of critical genes in the fim operon that encode for the uropathogenic virulence factor type 1 pili, we can better understand the functions of uropathogens and the pathogenesis of UTIs. This can advance antivirulence treatment versus broad-range antibiotics; in selectively targeting virulence factors, treatment becomes more effective. An in-depth understanding of UPEC virulence can aid in reducing UTI recurrence, antibiotic resistance, as well as associated healthcare costs.

### Conclusion

As shown in this study, we were able to identify 14 statistically significant (p<0.0001) binding motifs for putative fimS regulators in *E. coli* and *K. pneumoniae* reference strains. Among both species, PhoP was present in all strains, with the exception of one *K. pneumoniae* strain. As previous studies have shown that the deletion of phoP in avian *E. coli* has been shown to reduce biofilm formation,¹⁸ further studies may include the deletion of PhoP-binding motifs in human *E. coli* and *K. pneumoniae* strains to confirm whether it functions as an activator of colistin resistance and biofilm formation.

IHF (integration host factor) was also present among all *E. coli* reference strains and absent among all *K. pneumoniae* strains. In order to establish an association between *K. pneumoniae* as a pneumonia-causing pathogen and the absence of IHF as a fimS regulator, the deletion/mutation of IHF-binding motifs in *K. pneumoniae* sequences (clinical and reference) and any resulting effects on infection can be tested. Furthermore, CSTR and Hpr were present in all *K. pneumoniae* reference strains and none of the UPEC strains; their concurrent presence/absence could reveal the presence of a mechanism.

In addition, to confirm that every identified protein is a regulator of the intergenic region between fimE and fimA, and to determine a function/association with pathogenicity, an Electrophoretic Mobility Shift Assay could be conducted on the potential binding proteins.²⁰ Furthermore, the deletion and/or mutation of potential binding sites in uropathogenic strains can be analyzed on the basis of what functions they alter and how they affect type 1 pili production.

### Acknowledgements

This study’s success is due, in part to the work of the author’s Washington University Young Scientist Program (YSP) mentor Jesús B. Villicaña. He has been of significant support in providing guidance and genomic reference sequences for the project. Also, the author would like to thank her tutor, Bejan Bezhani, in supporting the research and guiding the format of the manuscript during the writing process. Furthermore, the author would like to thank of the YSP course instructors,
and the program director, Kitra Cates, for organizing the entire program and all of its resources, without which this project could not have been possible.

I would like to thank Abhishek Dev, a PhD student in Finance in Yale University, for mentoring and helping me throughout the research process. He helped me narrow down on the research topic and guided me on the various data analysis studies on the sample data. Without his constant guidance and support, this research paper would not have been possible.

Acknowledgments

References

Author

My name is Sarah Joseph, and I am a senior attending Parkway Central High School in St Louis, Missouri. I look forward to studying majors related to chemistry and biomedicine. As for my future career, I am largely undecided; however, bariatric, and internal medicine are of high interest to me.

DOI: 10.36838/v4i3.7
Replication Stress: Cancer's Double-Edged Sword

Keshav Saigal

American Embassy School, Chandragupta Marg, Chanakyapuri, New Delhi, Delhi 110021, India; keshavsaigal@yahoo.com

ABSTRACT: Cancer is fundamentally a disease of aberrant genetics. In contrast to nonmalignant cells, cancer cells are not only able to tolerate these genetic insults but are in fact enabled by them. While cancer treatment has progressed significantly in the last century, many current medical treatments rely on general cytotoxicity for rapidly dividing cells, rather than inhibiting the specific genetic lesions that enable malignant transformation. One promising area is the development of inhibitors targeted towards DNA replication. By targeting core pathways mutated in cancer, DNA replication inhibitor therapies offer new avenues for cancer treatment that spare patients the worst side effects of cytotoxic chemotherapy. One such target is RECQ1, a human helicase that shows great promise in cancer treatment when targeted as an adjuvant to other replication-based chemotherapies. In addition to summarizing the role replication stress plays in carcinogenesis, this paper discusses a triple combination of RECQ1, PARP1, and TOP1 inhibitors. While aggressive, this approach may increase the efficacy of these existing cancer therapies.

KEYWORDS: Biomedical and Health Sciences; Genetics and Molecular Biology of Disease; Cancer; Genomic Instability; DNA Repair; Replication Stress; Combined Inhibitors; RECQ1.

Introduction

Cancer is a broad grouping of diseases in which abnormal cells divide uncontrollably. Projected to kill over 600,000 people in the United States this year alone,¹ cancers are a subset of neoplasms – cells that grow without regulation and form a mass of tissue. However, in contrast to other neoplastic tissues, such as adenomas, cancer cells are malignant. They possess limitless replicative potential and may exhibit invasion into distant sites.

Despite their varying complexities and wide variety, many cancers may be characterized by certain key features, as laid out in 2000 by Hanahan and Weinberg’s The Hallmarks of Cancer and updated in Hallmarks of Cancer: The Next Generation.²,³ In the Hallmarks, Hanahan and Weinberg explicate six general characteristics of neoplastic disease — sustained growth signals, insensitivity to growth suppressors, avoidance of cell death, limitless replicative capabilities, induction of angiogenesis, and tissue invasion with the formation of metastases (Figure 1).² As cells progress towards a neoplastic state, they develop the abilities of each hallmark, capabilities that enable tumor growth and metastatic dissemination. In 2011, these hallmarks were updated to include reprogramming energy metabolism, avoiding immune damage, and acquisition of genomic instability, emphasizing the importance of continued investigation into this class of disease.³

While the Hallmarks deliver a descriptive analysis of cancer, inferential studies into the mechanisms underpinning the development of these hallmarks suggest an evolutionary model of carcinogenesis.²,³ The dominant theory of carcinogenesis today entails a hypermutator phenotype that allows for stochastically generated gene mutations to be selected in a manner similar to that observed in Darwinian evolutionary theory (Figure 2).⁵,⁶ It has been shown that many cancers achieve this hypermutator phenotype by developing genomic instability which allows cancer cells to sample a multitude of different phenotypes to best adapt to a microenvironmental niche.⁷ Greater knowledge of the mechanisms that produce this genomic instability will not only improve our understanding of cancer as a disease but also enable us to develop new, targeted treatments.

Discussion

At the core of genomic instability is deoxyribonucleic acid (DNA). In order for life to propagate, DNA must be replicated, thereby passing information from one generation to the next. In eukaryotic cells, this replication occurs once
during the cell cycle in the S-phase. This must be done with high precision, and there are multiple mechanisms that ensure this high-fidelity replication.

**DNA Structure:**

DNA is a double helix made of two strands of nucleotides linked by a sugar-phosphate backbone and annealed to one another by hydrogen bonding between complementary base pairs (Figure 3). The two strands run in opposite directions and are said to be antiparallel meaning one strand runs in the 5'→3' direction, while the other runs 3'→5'. The molecular structures of the nitrogenous bases comprising DNA allow complementary hydrogen bonds between specific bases — Adenine (A) with Thymine (T) and Guanine (G) with Cytosine (C).

The two antiparallel strands separate during DNA replication and each of the original strands serves as a template for a new strand. The new strands form when nucleotides are added by DNA polymerase (DNA pol) based on complementarity with the original strand. This results in semi-conservative DNA replication, in which two new DNA double helices, each consisting of one original strand and one daughter strand, are formed (Figure 4). This guarantees that each resulting daughter cell will have a full complement of double-stranded DNA.

**DNA Replication:**

While the principles of DNA replication are in many ways intuitive, the molecular process by which replication occurs involves a complex apparatus of proteins and nucleic acids that ensures proper regulation.

Before DNA replication can occur, the two original DNA strands must be separated to allow polymerases and other proteins to access the template strand. This separation is mediated by a group of enzymes called helicases, which couple their unwinding action to adenosine triphosphate (ATP) hydrolysis. As base-pair hydrogen bonds are broken and the DNA strands separated, ATP hydrolysis provides the energy for helicase translocation along the DNA double helix, typically in the 5'→3' direction.

This unwinding action produces a torsional strain on the DNA strands, which is rectified by a class of enzymes called the topoisomerases. This torsional strain results from DNA strands being twisted into tighter helices than their structures allow, termed positive supercoiling. This torsional strain can cause structural distortions and DNA damage. DNA gyrase, which is a member of the topoisomerase family, relaxes these twists via negative supercoiling — twisting against the helical conformation, unwinding, and straightening the structure of the helix. Single-stranded binding (SSB) proteins then bind to the single-stranded DNA (ssDNA) strands after they have been separated to prevent the strands from reannealing while also protecting the ssDNA from nucleolytic cleavage.

While the above enzymes are imperative for successful eukaryotic DNA replication, the key enzyme remains DNA polymerase, which catalyzes the addition of deoxyribonucleotide 5′-triphosphates (dNTPs) to the elongating DNA polymer, and which also possesses endonucleolytic activity. Free nucleotides align opposite their complementary base partners, and DNA polymerase moves along the two strands in a 5'→3' direction, covalently joining these free nucleotides.

Helicase's unwinding of the double-stranded DNA by breaking hydrogen bonds between base pairs occurs at specific regions called origins of replication. This action creates a replication fork, with the two-parent DNA strands running in antiparallel directions. While each DNA strand is replicated at a given replication fork, DNA polymerase can only synthesize in a 5'→3' direction, meaning one of the parent strands will be replicated continuously, and one will be replicated discontinuously. Consequently, the replication fork proceeds in a single direction by generating a continuous leading strand and a lagging strand consisting of discontinuous Okazaki fragments (Figure 5). Because DNA polymerase can only join nucleotides to the 3'-hydroxyl group of ribonucleotides, replication is initiated by de novo synthesis of an RNA primer by primase, thus providing DNA polymerase with a free 3'-hydroxyl to which a nucleotide may be joined. On the lagging strand, DNA polymerase elongates away from the replication fork, eventually meeting the RNA primer for the previously synthesized Okazaki fragment, prompting termination of its synthetic function and initiation of its endonucleolytic function. This results in degradation of the RNA primer of the previously synthesized...
Okazaki fragment and replacement of these ribonucleotides by DNA polymerase. The Okazaki fragments are then joined by DNA ligase to form a continuous strand by covalently joining the sugar-phosphate backbones together with a phosphodiester bond.

DNA replication must be highly accurate to maintain genetic integrity. High-fidelity DNA replication relies on three key aspects — specific nucleotide selection by replicative polymerases, exonucleolytic proofreading, and post-replication DNA mismatch repair (MMR).²⁹

As aforementioned, the eukaryotic genome is replicated by DNA polymerases, consisting of pol α, pol δ, and pol ε. Polymerase α synthesizes approximately 20–30 nucleotides during Okazaki fragment origination, and polymerase δ then extends this lagging strand.²⁰ DNA pol α lacks its proofreading; however, any errors it introduces are later removed by polymerase δ and its exonucleolytic proofreading activity.²¹,²² Errors that persist despite this proofreading are then corrected by MMR.²³ More complex errors must be corrected by the DNA damage response (DDR), discussed below.

DNA Repair:

Cells have multiple repair pathways in place to recognize and correct DNA damage, known collectively as the DNA damage response (DDR) system. There are five significant pathways — base excision repair (BER), homologous recombination (HR), non-homologous end joining (NHEJ), nucleotide excision repair (NER), and mismatch repair (MMR).²⁴ A brief description of BER, HR, and NHEJ follows, as these three are of significance to therapies discussed below.

The BER pathway is recruited when a molecular insult is made to a single nucleobase through exposure to radiation or genotoxic molecules.²⁵ This single-nucleobase damage is recognized by DNA glycosylase, which removes the damaged base from the strand by cleavage of the N-glycosidic bond via an activated water.²⁶ AP endonuclease then recognizes the abasic site (AP-site) and creates a nick in the phosphodiester backbone.²⁷ DNA polymerase then inserts a base at the AP-site, after which DNA ligase seals the nick.²⁸

Homologous recombination (HR) and non-homologous end-joining (NHEJ) both repair double-strand breaks in DNA.²⁶ Such breaks can form from exposure to reactive oxygen species (ROS) generated by normal respiratory metabolism inside a cell, external agents such as ionizing radiation, and certain chemotherapeutic drugs.²⁶ Additionally, during meiosis, HR facilitates recombination between homologous maternal and paternal chromosomes to produce gametes.²⁶

The primary mechanism by which HR occurs has been elucidated. Specialized helicases unwind the damaged DNA into its two constituent strands, allowing HR proteins to facilitate pairing between the damaged DNA and the same segment on the undamaged homologous chromosome.²⁶ DNA polymerases are then recruited to repair the break through replication, using the homologous strand as a template.²⁷ This process is relatively slow and requires extensive homology between regions of the two chromosomes, but results in the high-fidelity repair of damaged DNA (Figure 6).²⁷

![DNA repair pathways for double-strand breaks](ijhighschoolresearch.org)

In contrast, NHEJ requires no homology with another DNA duplex, instead of removing the damaged DNA and ligating the two break ends.²⁸ NHEJ is directed by microhomologies (short homologous DNA sequences) found on the ends of the double-strand breaks, as only when these sequences are compatible does NHEJ repair accurately.²⁸

DNA Repair Stress & Cancer Therapy:

DNA replication faces several obstacles, both intracellular and extracellular, which can cause replication stress wherein DNA synthesis is stalled or slowed, resulting in DNA damage.²⁹ In eukaryotic cells, replication origins form bidirectional replication forks.³⁰ Once DNA replication begins, the cells need to maintain accuracy, speed, and proper resource allocation to maintain polymerase processivity.³⁶ Thus, replication origins are fired in a regulated fashion, thereby preventing the potentially catastrophic consequences of replication stalling — cytotoxicity, deleterious mutation, and malignant transformation.³⁰

The primary mechanism by which fork stalling can cause DNA damage is through the formation of extended stretches of ssDNA, which occurs when replicative helicases continue to unwind the template DNA despite polymerase stalling.³¹ These stretches of ssDNA alongside the newly replicated, yet stalled, dsDNA activate a primer-template junction replication stress response.³² This results in the recruitment of replication stress response proteins, most notably the protein kinase ATM and Rad3-related (ATR).³³-³⁶ ATR is a major replication stress response kinase that phosphorylates substrates in order to help the cell continue DNA replication despite stress.³³-³⁶ ATR's activation inhibits cell cycle progression, gives more time for repair, and stabilizes the stalled fork.

ATR senses replication stress through interaction with protective ssDNA-binding proteins, such as ribonucleoprotein A (RPA), which bind to the extended ssDNA segments to prevent DNA damage.³⁷ Through this protein-protein interaction, ATR becomes active, setting off a chain of molecular signaling events that culminate in the phosphorylation of Chk1 kinase.³⁸ Additionally, ATR activation triggers the cell cycle S-M checkpoint, thereby preventing entry into mitosis when unreplicated DNA is present.³⁹ In the absence of ATR or Chk1, cells are unable to respond efficiently to stalled replication forks, resulting in the inappropriate firing of new replication forks, causing RPA depletion and DNA damage.⁴⁰ Furthermore, as the burden of ssDNA increases due to further
origin firing, there is a greater probability of replication fork collapse, resulting in DNA double-strand breaks (DSBs) and unreplicated DNA. If these cells are allowed to enter the M phase, their unreplicated chromosomes will activate cell death through mitotic catastrophe. Indeed, while the genomic instability conferred by increased replication stress can be advantageous to cancer, excessive replication stress can result in cell death, a phenomenon that holds great therapeutic potential.

While ATR responds to stalled replication forks specifically, there are several molecular perturbations may cause further stalling to occur. Nicks, gaps, and stretches of ssDNA, when encountered by DNA replication machinery, can be converted into DSBs, leading to catastrophic DNA damage. Moreover, DNA lesions — physical barriers to replication fork progression — generated by endogenous or exogenous DNA-damaging agents, such as UV light and chemical mutagens, can cause replication stress if left unrepaired. Furthermore, despite the high fidelity of DNA polymerases, both polymerase δ, and polymerase ε may fail to distinguish dNTPs from rNTPs. These mistakes should ideally be rectified by RNase H2-mediated ribonucleotide excision repair. However, these errors have been found recently to be made at an atypically high rate, producing yet another source of replication stress through the erroneous incorporation of the wrong nucleotide species. Additional sources of replication stress include collisions between replicative and transcriptional complexes, inherently unstable DNA regions, and drug-mediated nucleotide depletion.

As aforementioned, cancer is able to develop in many cases because of its genomic instability, which confers the potential for the rapid adaptation necessary for its unrestrained growth. Tumor cells achieve this instability through persistent replication stress, which is included as an “enabling characteristic” of cancer in the revisited Hallmarks review. However, this enabling characteristic has been exploited by therapeutics that target repair pathways needed to ameliorate this replication stress. This allows for the selective killing of cancer cells while sparing nonmalignant ones. Such therapeutics are based on the principle of synthetic lethality whereby one can kill cancer cells that have become “addicted” to a specific cellular pathway due to the absence of another, through its inhibition. Replication stress is essentially increased within these tumor cells, an exploitation of their predisposition for it. The most notable example of this strategy is the use of poly (ADP-ribose) polymerases (PARP) inhibitors in breast cancer-associated (BRCA1/2)-deficient cancers, discussed below.

Current Methods for Cancer Therapy Targeting DNA Replication Stress:

Camptothecin (CPT) – Topo I inhibitor. While topoisomerases can broadly be described as enzymes that cleave and repair phosphodiester bonds during normal DNA replication, there are actually two distinct subtypes: Topo I, which creates a single-stranded DNA nick, and Topo II, which creates a double-stranded DNA break. In addition to endonuclease activity, both Topo I and Topo II possess ligase activity, allowing them to repair the DNA breaks once the threat of damage from helicase-induced supercoiling is resolved. Because failure to repair these DNA breaks is lethal, this ligase activity can be targeted as a form of chemotherapy. Camptothecin (CPT) is a cytotoxic alkaloid that binds to Topo I, forming Topo I-cc, a stable covalent ternary complex with DNA. CPT exerts its cytotoxic effects in the S phase, during which it converts ssDNA breaks into 1-ended dsDNA breaks as the replication fork collides with Topo I-cc. Unlike 2-ended dsDNA breaks, which are caused by ionizing radiation, 1-ended dsDNA breaks cannot be repaired by NHEJ. As a result of Topo I-cc, Ataxia telangiectasia mutated (ATM) and ATR, both members of the phosphatidylinositol 3-kinase (PI 3-kinase) like family of protein kinases (PIKKs) involved in cellular responses to DNA damage, are activated, leading to recruitment of HR repair pathways. CPT additionally prompts hyper-phosphorylation of FANCJ, a DNA helicase important in the DNA replication stress response, and suppression of microsatellite instability in an ATR-dependent manner. Failure to activate any part of this stress response pathway in the presence of CPT results in DNA damage and apoptosis. Thus, CPT arrests cancer growth.

PARP1 Inhibitors. The poly (ADP-ribose) polymericases (PARP) are a family of DNA repair proteins that are important for BER (Figure 7). PARP1, a specific enzyme, detects sites of ssDNA and synthesizes poly (ADP) ribose (PAR), which then recruits other repair proteins to the damaged DNA site, leading to repair of the DNA damage through BER. When PARP, and therefore BER, is inhibited, SSBs accumulate and become DSBs. Accumulation of DSBs drives cellular dependence on repair via HR and NHEJ. Dependence on pathways such as HR, in which BRCA1 and BRCA2 proteins play a major role, has allowed inhibition of PARP for therapeutic benefit. Indeed, because cells with BRCA deficiencies are unable to repair DSBs, PARP inhibition has proven to be a potent method of chemotherapy in BRCA-deficient cancers. PARP inhibitors additionally enhance various other chemotherapies, such as irinotecan and carboplatin, by inhibiting the repair of SSBs induced by these agents. PARP inhibitors have been tested clinically, and there are currently several PARP inhibitors in development testing the two concepts — synthetic lethality in patients with deficient HR and removal of the cell’s ability to repair DNA damage caused by certain chemotherapies.
**RECQ1 as an Emerging Anticancer Target:**

Helicases are essential in maintaining genomic stability.¹¹ They play many roles in nucleic acid metabolism, including DNA replication, repair, transcription, and chromosome segregation.⁶⁴ Specifically, the RecQ family of helicases has proven indispensable for the maintenance of genomic integrity. These helicases serve as regulators of genetic recombination at stalled replication forks and help suppress elevated sister chromatid exchange.⁶⁵ RecQ Helicases are also involved in DSB repair, MMR, BER, and telomere maintenance.⁶⁵

One human RecQ helicase, RECQ1, has been recognized in recent years to play a particularly significant role. RECQ1 (also known as RECQL or RECQL1) is involved in multiple cellular processes, primarily implicated in the maintenance of chromosomal stability, and its loss has been linked to tumorigenesis.⁶⁵,⁶⁶ Studies of primary fibroblasts from RECQ1-knockout mice and human cells depleted of RECQ1 by RNA interference exhibited marked chromosomal instability.⁵⁷,⁶⁸ Cancer cells have marked genomic instability, and having to rely on fewer DNA repair pathways due to mutations in other DNA repair factors leaves them vulnerable to detrimental effects from disrupting a specific DNA repair pathway that plays a major role, such as RECQ1. Due to the functions in DDR and repair pathways that DNA helicases perform, targeted inhibition of RECQ1 has thus arisen as a novel approach that is informative for developing anticancer approaches harnessing DNA replication.⁶⁹ RECQ1 silencing offers a powerful emerging avenue for cancer treatment, as it offers a targeted therapy that exploits the replication stress inherent to cancer.

**Structure & Functions of RECQ1:**

The **RECQ** gene is located on chromosome 12p12 and it encodes a 649 amino acid protein weighing 73 kDa.⁷⁰ Like replicative helicases, RECQ1 catalyzes ATP-dependent unwinding of DNA and can unwind a plethora of DNA structures, including stalled replication forks.⁷¹ It is the most abundant RecQ homolog present in humans,⁷² and contains four domains: N-terminus, a core helicase, the RecQ-specific C-terminal (RQC) domain, and C-terminus (Figure 8). In addition to helicase activity, RECQ1 also activates the annealing of complementary ssDNA in an ATP-independent manner, facilitating HR DNA repair.⁷³

**Capture Figure:**

**Figure 8:** Human RECQ1 Protein; A – Crystal Structure; B – Domain Structure.⁷⁴

**RECQ1 & Cancer Expression:**

Overexpression of RECQ1 has been observed in multiple malignancies,⁷⁵ including glioblastoma, multiple myeloma, ovarian cancers, hematological cancers, hepatocellular carcinoma, and head and neck cancers.⁷⁶-⁸¹ Silencing RECQ1 results in cancer cell death, suggesting its activity is cancer-protective.³² Furthermore, depletion of RECQ1 sensitizes cancer cells to several genotoxic chemotherapeutics, including camptothecin, hydroxyurea, and temozolomide.⑧ This finding indicates that RECQ1 may be necessary for cancer resistance to these agents. Moreover, RECQ1 expression is increased in response to DNA damage, enabling some malignant cells to persist despite drug-mediated replication fork stalling.⁷⁴ This enables some cancer cells to survive the cytotoxic effect of chemotherapeutic drugs, and therefore, additional inhibition of RECQ1 could be effective.

**RECQ1 Silencing in Breast & Brain Cancer:**

Evidence for such efficacy has been shown with mutations in RECQ1 and subsequent implications in breast cancer pathogenesis. A clear connection between RECQ1 and estrogen receptor alpha (ERα) has been established.⁸⁴ ERα is a master regulatory transcription factor in breast cancer, and as reported by Lu et al., the expression of ESRI, the gene encoding ERα, is directly activated by RECQ1.⁸⁴ They report that >35% of RECQ1 binding sites were also bound by ERα genome-wide, suggesting that RECQ1 may act as a novel cofactor for ER-alpha. These findings demonstrate a new mechanism by which RECQ1 regulates gene expression in ER-positive breast cancer cells, thereby driving disease and serving as a critical elaboration on several previous reports suggesting RECQ1 as a breast cancer susceptibility gene.⁷⁴

Immunohistochemical and Western blot analyses comparing malignant and nonmalignant cells have also implicated RECQ1 as a driver of glioblastoma.⁷⁶ Mendoza-Maldonado et al. demonstrate that RECQ1 is overexpressed in tumors, dramatically so in human brain glioblastoma relative to control brain tissues.⁷⁶ They demonstrate higher levels of DNA damage in RECQ1-depleted glioblastoma cells, indicating that RECQ1 plays a protective role in glioblastoma survival, as well as a significant decrease in T98G proliferation when RECQ1 is depleted.⁷⁶ Furthermore, this work also shows that silencing RECQ1 expression sensitizes glioblastoma to hydroxyurea and temozolomide, two agents widely used to treat brain tumors — as RECQ1 is essential to genomic stability, DNA damage was increased and thus T98G and U87 cells were more susceptible to treatment.⁷⁶

Moreover, using xenotransplantation of human glioblastoma cells into zebrafish embryos, Vittori et al. were able to elaborate on the findings of Mendoza-Maldonado et al.⁸⁵ They confirmed the growth-inhibiting effects of RECQ1 silencing in U87 glioblastoma cells in vivo in a zebrafish embryo model.⁸⁵ Using this approach, they were able to corroborate the in vitro effects of RECQ1 silencing observed by Mendoza-Maldonado et al. in an in vivo setting.⁸⁵ The results, by looking at fluorescently-labeled cells and observing tumor-associated angiogenesis, effects of drugs, and involvement of signaling pathways in tumor progression

**Image References:**

A – Crystal Structure; B – Domain Structure.
demonstrated that RECQ1 silencing reduces the rate of increase in the number of U87 tumor cells through the result of cell cycle perturbation by RECQ1 knockout.⁸⁶-⁸⁹

Looking Forward:

With their work in breast and brain cancer, Lu et al., Vittori et al., and Mendoza-Maldonado et al. solidify the efficacy and role of RECQ1 inhibitors in cancer therapy.⁷⁶,⁸⁴,⁸⁵ However, while sole RECQ1 inhibition has potential, when combined with the aforementioned TOP1 and PARP1 inhibitors, it may have a greater impact. In research recently published in Nature Structural & Molecular Biology by Berti et al., it was demonstrated how cancer cells respond to the damage caused by topoisomerase I inhibitors.⁹⁰ When TOP1 inhibition causes replication stress via the formation of nicks, the replication machinery pauses and reverses its course instead of colliding with the TOP1-induced DNA lesion. They find that two important cellular proteins, PARP and RECQ1, regulate this fork reversal mechanism.⁹⁰ They demonstrate an interplay between these two proteins whereby PARP activity at sites of CPT-induced DNA damage inhibits RECQ1-mediated premature replication fork restart.⁹⁰ This interplay suggests a potential therapeutic benefit in targeting RECQ1 and PARP as adjuvants to TOP1 inhibition, thus preventing repair of CPT-induced DNA damage by PARP. With TOP1 and PARP1 inhibition, any disruption of RECQ1 activity would prevent cells from restarting reversed forks. This proposed triple combination treatment highlights how RECQ1 inhibitors with PARP1 inhibitors like Olaparib and TOP1 inhibitors like Camptothecin, could cause excessive replication stress in malignant cells, resulting in targeted cell death. Cancer cells relying on DNA repair pathways like TOP1 and PARP1 would not even be able to turn to RECQ1 if these were inhibited. Evidence for such treatment has arisen in work surrounding multiple myeloma cells (MMCs) — Viziteu et al. demonstrated RECQ1 depletion sensitized MMCs to PARP inhibitors.⁷⁷ RECQ1 inhibitors involved like so, as adjuvants to others harnessing replication stress, should ideally be the new target in battling cancer as they could increase the efficacy of combinational cancer therapies.

However, RECQ1-based therapies in cancer are still mainly at a preclinical stage due to challenges in applying helicase inhibitors. In general, it is difficult to develop highly specific small molecule inhibitors. A significant barrier is bioavailability — the extent and rate at which the drug can access the site of action by entering the systemic circulation. Brosh & Datta recognize that among other cellular bioavailability parameters, inhibitor aqueous solubility, nonspecific binding to the cell membrane and extracellular matrix, cellular uptake, and intracellular metabolic stability must be considered for successful exploitation.⁹¹ With helicases specifically, another challenge lies in ensuring helicase-interacting compounds reach their desired subcellular location — one potential method for nuclear-targeted delivery is nanoparticles.⁹¹ With nuclear localization sequences (NLS), they should help deliver the small molecule inhibitors to target tumor sites. This specific delivery heightens the therapeutic effects and lessens negative consequences if achieved with sufficient concentration. With issues on the potency of helicase inhibitors, they need to be able to have their intended biological effects at minimal concentrations. Since the field of helicase inhibitors is relatively new, there are concerns with drug resistance mechanisms as well. While the basic principle is to compromise a helicase-dependent pathway of repair leading to synthetic lethality, other repair pathways might get involved that interfere with the helicase inhibitor.⁹¹ This requires further research and prioritized attention. Lastly, there is a challenge with the structural configuration of inhibiting DNA repair proteins. The inhibitor compound needs to approach its target effectively; however, the proteins’ flexibility and intrinsic conformational states cannot ensure a good fit.⁹¹ Brosh & Datta report that machine learning algorithms and artificial intelligence can help such treatments with advanced structure-based drug design.⁹¹ These obstacles need to be addressed if targets for therapy from cancer predisposition syndromes associated with mutations in DNA repair genes, like helicase inhibitors, can be built.

■ Conclusion

In summary, although replication fork instability can be a boon for cancer adaptation, we can leverage our knowledge to turn this stress into an advantage, exploiting its vulnerability. While RECQ1 is one such promising avenue, the complexity of replication makes it rich with therapeutic potential. Just as recognition of immune evasion as a hallmark of cancer has enabled the development of novel therapies in the form of checkpoint inhibitors and CAR T-cells, so too can the recognition of genomic instability as a central feature of cancer development lead to groundbreaking targeted cancer treatments.²,³

■ Acknowledgements

I would like to thank Paul Young (Tri-Institutional MD-PhD Program) for his mentorship, guidance, help, and support throughout the research process and with the paper.

■ References

70. Seki, M. Molecular cloning of cDNA encoding human DNA helicase Q1 which has homology to Escherichia coli Rec Q helicase and localization of the gene at chromosome 12p12. Nucleic Acids Research 1994, 22 (22), 4566–4573.
76. Mendoza-Maldonado, R. The human RecQ1 helicase is highly expressed in glioblastoma and plays an important role in tumor cell proliferation. Molecular Cancer 2011, 10 (1), 83.
78. Sanada, S. RecQ1 DNA Repair Helicase: A Potential Therapeutic Target and a Proliferative Marker against Ovarian Cancer. PLoS ONE 2013, 8 (8).

Author
Keshav Sagiul, a senior at the American Embassy School, is fascinated by the biological sciences and plans to major in molecular and cellular biology. A trained EMT, he hopes to continue into the medical field while still researching cancer. Keshav enjoys playing basketball and chess in his free time.
Differentially Expressed Genes and Pathways Involved in the Development of PTSD in World Trade Center Responders

Kevin Toledo, Harmanpreet Kaur, Miguel A. Javiel, Jennifer Barreto
Medical Academy for Science and Technology at Homestead, 1220 NW 1st Ave, Homestead, FL 33030, USA; ktoledo0456@gmail.com

ABSTRACT: Post-traumatic stress disorder (PTSD) is a chronic psychological disorder that develops due to exposure to environmental trauma, resulting in debilitating impacts on cognitive function, emotions, and psychological recovery. The differential expression of key genes in the human genome is involved with the progression of PTSD and contributes to PTSD-related symptoms. Epigenetic alterations control the expression levels of specific genes by altering DNA methylation and histone acetylation patterns on them. Since genetic expression markers have been linked with the development of PTSD, our study aims to 1) confirm previous genetic expression markers, 2) identify new genetic expression markers, and 3) identify the biological and pathological pathways that are involved in the onset and progression of PTSD. Using the Gene Expression Omnibus (GEO) database provided by National Center for Biotechnology Information (NCBI), we retrieved a sample of 324 World Trade Center responders, who were sorted into three cohorts: PTSD patients, former PTSD patients, and non-PTSD patients. Using RNA sequencing technology, transcriptome-wide RNA expression patterns were determined for each participant. Genes were analyzed using DESeq2 to outline a variety of genetic expression markers that are prime targets for PTSD symptoms. Additionally, neurological, immune, and metabolic pathways were explored to reveal the associations between PTSD and other comorbid disorders.

KEYWORDS: Computational Genomics; PTSD; Gene expression; Transcriptome; Epigenetics; RNA-sequencing; Biomarkers.

Introduction
Post-traumatic stress disorder (PTSD) is a disabling mental condition characterized by flashbacks, avoidance, and hyperarousal symptoms that produce adverse changes in cognition, mood, and personal functioning as a response to a traumatic event.¹ Although 60-90 percent of people are exposed to at least one traumatic incident during their lives,² only 3.9 percent of US population has a PTSD diagnosis.³ The psychological trauma associated with recurring images of past traumatic or life-threatening events leads to flashbacks, nightmares, insomnia, and intrusive thoughts for many PTSD patients, often over extended periods of time. PTSD is linked with considerable morbidity, and patients with PTSD have a greater risk of suicidal conduct than the general population.⁴ PTSD is also highly associated with other mental disorders such as depression, drug addiction, and panic disorder. These disorders place an even greater burden on a patient's emotional and physical well-being. Contemporary methods of intervention include several psychological treatments, predominantly eye movement desensitization and reprocessing (EMDR) and trauma-focused cognitive behavioral therapy (TF-CBT).⁵

Despite PTSD becoming one of the most diagnosed disorders among mental health professionals, the full scope of its prevalence is not understood.⁶ Since PTSD is directly impacted by environmental trauma factors, the function of non-genetic variables in the condition has been of interest. Various studies have shown that following differentiation, cells exhibit altered gene expression patterns, which are transferred during cell division to the daughter cells. This concept makes up the emerging field of epigenetics, which involves the effect of heritable mechanisms on gene expression, the changes of these factors throughout development, and the transmission of these factors to the offspring.⁷ In all of these scenarios, the genome and genetic information are not being altered; rather, the expression of these genes are being controlled by epigenetic factors, namely DNA, RNA, proteins, and chemical groups.

One of the most thoroughly studied mechanisms of epigenetic inheritance is DNA methylation. Put simply, methylation of a specific part of the genome is associated with gene silencing. Methylation occurs on silent genes and functions as a cellular lock, preventing transcriptional factors from binding to the promoter region of these genes, thereby influencing gene silencing. Studies have shown that the patterns of methylation are influenced by the DNA sequence itself, by how tightly packed the DNA strands are, and by other epigenetic factors which all coalesce to form a complex mechanism of gene expression.⁸

Studies have confirmed the connection between genetic expression patterns and the pathophysiological development of PTSD. Specifically, PTSD appears to develop from a combination of inherited genetic and epigenetic risks as well as acquired epigenetic marks that result from childhood and adult trauma exposure.⁹ To better understand the etiology of PTSD, we need to understand the role of environmental influences, genetic interactions, allele types, and epigenetic modifications. The current study utilizes a promising
methodology to explain the gene expression differences underpinning PTSD, a transcriptome-wide design.

Many of the current transcriptome-wide gene studies that have been conducted on PTSD have generally relied on small sample sizes (N < 40). Notably, Mehta et al.¹⁰ used a 169-participant sample of participants who were abused as children with 61 PTSD cases, and attributed differential gene expression to specific DNA methylation patterns that developed as a result of the maltreatment. Tylee et al.¹¹ analyzed 50 PTSD and control samples and using the 23 full-length transcripts was able to predict PTSD with a 70% accuracy. Logue et al.¹² studied 115 PTSD participants and 28 controls and identified 41 differentially expressed, though only 1 gene remained after multiple testing correction. Bann et al.¹³ utilized RNA-seq, miRNA, and DNA methylation data and found an inverse relationship between differential expression and miRNAs; the study also concluded that inflammation in PTSD may be the result of epigenetic factors. Even studies with larger sample sizes (>40) that have investigated key candidate genes related to PTSD have produced inconsistent results, which may be a result of a lack of statistical power or methodological errors.¹⁴

The current study aims to expand the current understanding of how PTSD develops using a large and relatively homogeneous sample. We made use of a publicly available RNA sequencing dataset of 324 World Trade Center responders. Previous studies on PTSD and gene expression patterns have included PTSD and non-PTSD cohorts, but no study to date has included a “past” cohort in which participants were previously diagnosed with PTSD but recovered. A past cohort would allow us to gain insight into what it means to recover from PTSD as well as understand the gene expression changes that occur when a person overcomes PTSD. In this study, the past cohort (N = 42) will be compared to PTSD and non-PTSD cohorts to allow us to better understand the transcriptomic changes that occur during PTSD recovery. This research will not only contribute to our current understanding of how PTSD develops, but it will also highlight the target genes and pathways that are driving its development.

## Methodology

The participant gene expression data utilized in this study was collected by Dr. Kuan of the University of Wisconsin and her colleagues.¹⁵ The researchers recruited a total of 324 participants from the Stony Brook World Trade Center Health Program: 201 subjects did not have PTSD, 81 had PTSD at the time of the study, and 42 previously had PTSD but did not at the time of data collection. All participants were World Trade Center responders who experienced a single major traumatic event – that is, the World Trade Center terrorist attack on September 11, 2001. This constraint limited the heterogeneity of the environmental exposure for PTSD development in the sample. To ensure that differential expression patterns were primarily due to PTSD, females were excluded from the study as females express significantly different gene expression patterns from males.¹⁶ To categorize the participants into the past, never, and current groups, the researchers used the Posttraumatic Stress Disorder Checklist (PCL-C) developed by psychology professor Frank Weathers¹⁷ that assesses the severity of the symptoms experienced by the participants related to the trauma.

### Table 1: Demographics of all three cohorts depicting age and race clinical data.

<table>
<thead>
<tr>
<th>Clinical Characteristic</th>
<th>Never (N = 201)</th>
<th>Current (N = 81)</th>
<th>Past (N = 42)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>51.36</td>
<td>51.57</td>
<td>51.36</td>
<td>0.331</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>0.20</td>
<td>0.76</td>
<td>0.26</td>
<td></td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caucasian</td>
<td>69 (85.2%)</td>
<td>33 (78.6%)</td>
<td>181 (90.0%)</td>
<td>0.100</td>
</tr>
<tr>
<td>Other</td>
<td>12 (14.8%)</td>
<td>9 (21.4%)</td>
<td>20 (10.0%)</td>
<td></td>
</tr>
</tbody>
</table>

Blood tissue was collected by Dr. Kuan and her colleagues, which was later profiled using RNA sequencing (RNA-seq) at the Roswell Park Cancer Institute Genomic Shared. Specifically, Dr. Kuan’s team conducted transcriptome-wide expression analysis using RNA sequencing for the 324 participants. RNA sequencing is used to detect the transcriptome of RNA in each sample.¹⁸ RNA was isolated using the Paxgene Blood Collection kit (Qiagen) and purified. Abundant amounts of mRNA were removed from the sample using the rRNA depletion process.¹⁹ After the ribosomal RNA was removed, the remaining RNA was fragmented to allow for measurable sizes in the sequencer. RNA was then converted into complementary DNA (cDNA) to allow for a more stable form of molecule that can be easily modified and amplified as needed.²⁰ Multiple indexing adapters were ligated to the ends of the cDNA fragments to allow for the fragments to hybridize onto the flow cell and so the sequencing machine can recognize the fragment pieces.²¹ The fragments are then amplified, and added onto a HiSeq Flow Cell, and then sequenced using a HiSeq2500 Sequencer (Illumina Inc., San Diego, CA).

We retrieved the raw data from Kuan et al.¹⁵ via the Gene Expression Omnibus (GEO) database provided by the National Center for Biotechnology Information.²² The raw data is freely accessible with accession number GSE97356. In order to analyze the data, we used the DESeq2 software²³ provided via the R programming language. Using the DESeq2 library, we normalized the data, accounting for differences in terms of library size in the collection of the samples (also known as a sequencing depth) and RNA composition. In order to allow the program to run faster and eliminate some genes that were highly unlikely to be differentially expressed, all the lowly expressed genes (<10 transcripts) were excluded from the dataset. Out of the original 25,830 genes accounted for, 24,547 remained after the manual exclusion of these lowly expressed genes. The DESeq2 library, after calculating dispersion estimates and fitting dispersion models to the data, excluded outliers that are unlikely to be differentially expressed and adjusted dispersion estimates based on its algorithm to minimize the likelihood of false positives. Once normalized, we used DESeq2 to complete a differential gene expression analysis.
analysis and create generalized linear models (GLMs) based on a negative binomial distribution for each gene. To test for statistical significance, we performed a Wald test, calculated by dividing the log fold change (LFC) by its standard error.²³ This results in a z-statistic that can then be compared to a normal distribution to obtain a p-value for each gene.

Since we used a p-value of less than 0.05, with a dataset of over 20,000 genes, we can expect over 1,000 false-positive results. This is known as the multiple testing problem and it arises from the fact that for each of the 25,830 genes, a p-value was calculated. Thus, we used a multiple test correction technique Benjamini–Hochberg correction²⁴ to reduce the false discovery rate (FDR). With FDR < 0.05, we should expect less than 5% of those genes found to be differentially expressed after adjustment to be false positives. Therefore, we determined differential gene expression for those genes whose adjusted p-value was less than 0.05.

The DAVID Bioinformatics Resource²⁵ was used to perform a KEGG analysis of the differentially expressed genes in the comparison between the never and current groups. KEGG is a set of databases containing information about genomes, biological pathways, and diseases that report the pathways likely to be involved in the observed differential gene expression. Using this platform, probable correlations between the differentially expressed genes and certain diseases and biological pathways were determined.

Results

After conducting the Benjamini–Hochberg correction, we made separate comparisons between the three groups (never, current, and past). When comparing the never and current groups, a total of 1051 genes were differentially expressed, with 516 genes being under expressed and 535 genes being over expressed in subjects with PTSD. The volcano plot (Figure 1) depicts participant-wide expression patterns by showing the relationship between the log fold change and -log₁₀ of the adjusted p-value after multiple testing corrections. Each point on the volcano plot is representative of a gene, with differentially expressed genes (FDR < 0.05) at the top in red. A total of 238 genes displayed an LFC > 1.1 and 309 genes an LFC < 0.9. A total of 53 genes showed an LFC > 1.2 and 36 genes an LFC < 0.8.

![Volcano plot depicting all differentially expressed genes based on their log2 fold change (LFC) and their transformed FDR. Genes who remained differentially expressed after multiple testing corrections (FDR < 0.05) are depicted in red.](image)

Our data has confirmed previously identified genes with a direct correlation to PTSD, such as FKBP5,¹² SOX5, NOS1AP,²⁷ C9orf84 (SHOC1), and NFKB1.²⁸ Violin plots with embedded box plots are displayed in Figure 2 for each of these genes, with the normalized count being depicted for all samples in the current and never groups. It should be noted that extreme outlier values were filtered out of the plots to make the violin plots easier to observe.

![Violin plots with embedded box plots displaying the normalized counts in the current and never groups for a variety of genes that have been previously identified in the literature.](image)

A heat map is depicted in Figure 3 that compares the 283 samples for the never (N = 201) and current group (N = 81) to the DE genes with average expression mean > 50, FDR < 0.05, and log₂ FC > 0.3 (Figure 3). The color gradient identifies the z-statistic between that gene and the sample, with the extrema representing higher differential expression. The scaling occurred across the rows, and therefore no meaningful comparisons can be made between genes, but instead between samples.

When comparing the past and never groups, 138 genes were differentially expressed but only 3 remained significant after correction (and a few others with a p-value slightly above 0.05). As for the past and current groups, there was no differential gene expression, and the gene with the lowest FDR...
has an FDR of 0.16. For a more detailed discussion of these groups, please see the Discussion section.

The KEGG analysis results from the DAVID resource are depicted in Table 2.

Table 2: Pathways identified by KEGG pathway analysis. For each pathway, the number of genes differentially expressed are shown as well as the adjusted p-value.

<table>
<thead>
<tr>
<th>KEGG pathway</th>
<th>Gene Count (percent of total)</th>
<th>Benjamini Adjusted P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Systemic lupus erythematosus</td>
<td>29 (2.8%)</td>
<td>3.1E-4</td>
</tr>
<tr>
<td>Alcoholism</td>
<td>33 (3.2%)</td>
<td>3.8E-4</td>
</tr>
<tr>
<td>Huntington's disease</td>
<td>24 (2.3%)</td>
<td>7.6E-3</td>
</tr>
<tr>
<td>Oxidative phosphorylation</td>
<td>19 (1.9%)</td>
<td>7.0E-3</td>
</tr>
<tr>
<td>Non-alcoholic fatty liver disease (NAFLD)</td>
<td>19 (1.9%)</td>
<td>3.0E-3</td>
</tr>
<tr>
<td>Alzheimer's disease</td>
<td>20 (2.0%)</td>
<td>3.6E-3</td>
</tr>
<tr>
<td>Parkinson's disease</td>
<td>17 (1.7%)</td>
<td>8.0E-3</td>
</tr>
<tr>
<td>Ubiquitin mediated proteolysis</td>
<td>16 (1.6%)</td>
<td>1.3E-3</td>
</tr>
<tr>
<td>Neurotrophin signaling pathway</td>
<td>12 (1.2%)</td>
<td>7.0E-4</td>
</tr>
<tr>
<td>HIF-1 signaling pathway</td>
<td>10 (1.0%)</td>
<td>8.6E-4</td>
</tr>
<tr>
<td>Pymatin metabolism</td>
<td>10 (1.0%)</td>
<td>9.0E-4</td>
</tr>
<tr>
<td>ErbB signaling pathway</td>
<td>9 (0.9%)</td>
<td>9.0E-5</td>
</tr>
<tr>
<td>NF-kappa B signaling pathway</td>
<td>9 (0.9%)</td>
<td>9.0E-5</td>
</tr>
</tbody>
</table>

Confirmatory Results:
Our findings confirmed a variety of genes and biological pathways that have been previously identified. Below, these confirmed pathways are listed with respective references to the studies that have identified them.

Table 3: Table depicting the confirmatory results for previously identified genes or pathways of interest supported by our findings. The main gene (or genes) of interest are shown, as well as the fold change they expressed in our study and the accompanying literature for further reading.

<table>
<thead>
<tr>
<th>Pathway</th>
<th>Gene of Interest</th>
<th>Fold Change</th>
<th>Literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glucocorticoid Pathway</td>
<td>FKBP5</td>
<td>1.27</td>
<td>Szeszko et al. 13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Kang et al. 14</td>
</tr>
<tr>
<td>NF-κB Pathway</td>
<td>NFKB1</td>
<td>1.05</td>
<td>O'Donovan et al. 17</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Guardado et al. 18</td>
</tr>
<tr>
<td>Systemic Lupus</td>
<td>Histone-coding</td>
<td></td>
<td>Roberts et al. 19</td>
</tr>
<tr>
<td>Erythematosus</td>
<td>Genes</td>
<td></td>
<td>Case et al. 20</td>
</tr>
</tbody>
</table>

Neurotrophin Signaling Pathway:
The KEGG analysis concluded that 12 differentially expressed genes within our data were associated with the neurotrophic signaling pathway. This pathway is responsible for neuroplasticity and maintaining the integrity of the hippocampus.12 Given that this pathway is involved in moderating fear responses as well as the physiological response to stress, its downregulation exhibited in the PTSD group (with the majority of genes in the pathway being under expressed in PTSD patients) appears to contribute to the symptoms experienced by those affected by PTSD.

DOI: 10.36838/v4i3.9


**ErbB Signaling Pathway:**

Pathway analysis found 9 differentially expressed genes within our data that are components of the ErbB signaling pathway. Among these genes is the MTOR gene, which codes for the mTOR protein, which plays a key role in the cellular response to stressors. This gene (fold change = 1.06) plays a key role in the development of fear memories. Fifield et al. concluded that the mTOR pathway was associated with the formation of associative and non-associative fear memories, with its activation contributing to the development of anxiety disorders.³⁴ Chen et al. also confirmed these results by concluding that the ErbB4 gene (a part of the ErbB signaling pathway) was responsible for fear conditioning, and the inhibition of this gene resulted in impaired fear conditioning function in mice.³⁵ The SYT1 gene and its interactions with Wnt pathways modulates subsequent neurotransmitter release in synapses.³⁸ More recently, mice models have shown that increased expression levels of the SYT1 gene may be due to the hyperacetylation of H3K9 and H3K14, which are promoter regions for SYT1.³⁹

**Wnt Signaling Pathway:**

Previous studies demonstrated that the development of PTSD may be mediated by the Wnt signaling pathway due to malfunctions in synaptic transmission in the brain that are associated with trauma development.³⁶ This pathway plays a critical role in regulating synaptic plasticity and memory formation.³⁷ We identified the SYT1 gene as a possible genetic expression marker as it was upregulated in the PTSD cohort with a fold change of 1.26. The upregulation of the SYT1 gene and its interactions with Wnt pathways modulates subsequent neurotransmitter release in synapses.³⁸ More recently, mice models have shown that increased expression levels of the SYT1 gene may be due to the hyperacetylation of H3K9 and H3K14, which are promoter regions for SYT1.³⁹

**Major Depressive Disorder:**

Previous studies have noted overlapping gene expression patterns between PTSD and major depressive disorder (MDD) patients. Our data showed that the DOCK10 gene showed around an 8% increase in expression in the PTSD cohort. Similarly, the GLS and CD47 genes were differentially expressed, though after Benjamini–Hochberg correction, there was no statistical significance (with adjusted p-values of 0.09 and 0.19, respectively). These three genes were categorized as relevant biomarkers for tracking and predicting depression.⁴⁰ The SOX5 gene (previously associated with mood disorders) was also differentially expressed between the current and never cohorts, with expression levels 30% higher in the PTSD group (FDR < 0.05). Lastly, our results identified the C9orf84 gene (a genetic expression marker for PTSD as it displayed a fold change of 1.31. This gene has been linked to MDD pathways⁴⁴ and is therefore a gene of interest for the development of PTSD.

**Schizophrenia:**

Our study also found ties between PTSD and schizophrenia. Seow et al. indicated that PTSD and schizophrenia share neurological risk factors and superficial symptoms.⁴² The NOS1AP gene was identified as overexpressed by a fold change of 1.61. Overexpression of NOS1AP was found in postmortem samples of schizophrenia patients.⁴³ The protein this gene encodes binds to synapsin and disrupts the signal transduction of the N-methyl d-aspartate receptor complex, which plays a role in memory formation.⁴⁴ The G allele of this gene has been associated with PTSD and severe depression.²⁷ Our findings substantiate the connection between schizophrenia and PTSD with relation to the NOS1AP gene.

**Parkinson's Disease:**

The KEGG analysis found 17 differentially expressed genes within the PTSD cohort that are associated with Parkinson's disease (PD). There is a comorbid association between PTSD and PD but the exact biological mechanisms behind the association are yet unknown.⁴⁵ We identified the TRIM11 gene as a possible genetic expression marker with a fold change of 0.93. TRIM11 is a protein known to restore cell viability as it prevents the formation of protein aggregates and deposits that interact with the TRIM SUGO ligase to degrade proteins that are malfunctioning.⁴⁶ The downregulation of this gene is likely linked with neurodegeneration and may explain the connection between PTSD and PD.

**Alzheimer's Disease:**

Studies have shown that individuals with PTSD have a greater chance of developing Alzheimer's disease (AD).⁴⁷ Our analysis of KEGG pathways found 20 genes that were implicated in the progression of AD. Specifically, the NDUF1 gene was significantly downregulated in the PTSD cohort with a fold change of 0.81. The connection is likely due to a missense mutation which results in significant neurodegenerative side effects.⁴⁸ The hypermethylation of the NDUF1 sequence has been linked to under expression of the gene, indicating that epigenetic factors may be involved in this gene's downregulation.⁴⁹ Given the decreased expression of the NDUF1 gene in PTSD patients, and the correlation of the gene to the development of neurodegenerative disorders, the gene should be further researched for possible associations with PTSD development.

**Metabolic Pathways:**

Non-alcoholic fatty liver disease (NAFLD) is a condition associated with excessive fat storage in the liver.⁵⁰ NAFLD is a comorbid condition with PTSD and other mental disorders, due to the common pathogenesis of oxidative stress pathways and lack of mitochondrial function.⁵¹ Individuals with mitochondrial dysfunction have the capacity to execute needed daily life needs but have deficient mitochondrial capacity to respond to traumatic exposure.⁵² The downregulation of the NDUF1 gene (fold change = 0.81) has been associated with the development of reactive oxygen species (ROS), which is generated by oxidative stress within cells. The formation of ROS is linked to the development of NAFLD.⁵³ Therefore, the pathological connection of 19 DE genes with the NAFLD pathway from our KEGG analysis seems to derive from the differential expression of genes linked to oxidative stress.

**Future Studies and Limitations:**

Since the expression patterns of these samples were of blood tissue, predicting neurological associations in brain tissue is a possible limitation of our paper. However, studies have noted that blood expression patterns are accurate predictors of gene
expression in the brain. For instance, blood gene expression patterns were highly predictive (with 85 to 90 percent accuracy) of neurodegenerative gene expression patterns in the brain as well as disease progression.⁵⁴ This underscores the importance of studying blood expression patterns, since studying brain tissue would require deceased participants. Nevertheless, it is important to note this limitation as differences in genetic expression in other tissues besides blood may be present. Future studies should aim to study other tissues as well as verify the validity of using blood in understanding the development of PTSD. Another limitation of our work is that all participants were males and many of them were of the Caucasian ethnicity. Therefore, these findings may not generalize to females or other ethnicities, and future research needs to be done for these populations.

While studies have shown that psychotherapy can reverse a few epigenetic modifications made after trauma exposure⁵⁵ and help PTSD patients cope with their symptoms, these behavioral changes are not permanent and some patients experience a relapse of symptoms after treatment, often resulting in increased alcoholism and drug intake.⁵⁶ Since epigenetic marks were the cause of the altered expression levels that resulted in phenotypic differences in PTSD patients, one could theoretically reverse these marks to alleviate the symptoms. However, the field of epigenetics is in its infancy and the possibility of modifying the epigenome is but a mere contemplation. Thus, future research should begin considering whether modifying the epigenome would be a plausible and practical treatment.

■ Conclusion

In this paper, we have outlined a variety of gene expression markers that are potential prime targets for epigenetic alterations. We have also presented a variety of neurological, immune, and metabolic pathways that are associated with PTSD and other stress-related and depressive disorders. Undoubtedly, other pathways allowing for PTSD progression may exist, especially considering the complex and highly interconnected nature of this disorder. In the immediate future, it is crucial to continue validating (or challenging) the existing literature and the biomarkers and pathways identified. This will allow for future treatments to be highly calibrated and precise to the specific regions of interest and for diagnostic and screening tools to detect abnormalities before the onset of severe symptoms.

■ Acknowledgements

We would like to thank our research mentor Jennifer Barreto as well as the Summer Research Internship at Florida International University for overseeing this research. We would also like to thank Dr. Sana Nasim from Harvard Medical School for reviewing this paper.

■ References


50. Maurice, J.; Manousou, P. Non-alcoholic fatty liver disease. *Clinical medicine*. 2018, 18(3), 245-250. https://doi.org/10.7861/clinmedicine.18-3-245


**Author**

Kevin Toledo is a junior at the Medical Academy for Science and Technology in Homestead, Florida. He is interested in majoring in Neuroscience and Computer Science in college and pursuing an MD-PhD program. His research interests involve molecular medicine, neurology, and genomics.

Miguel Javiel is a Junior at Medical Academy for Science and Technology in Homestead, Florida. He is interested in majoring in Biomedical Engineering in college and pursuing an MD-PhD program. His research interests involve molecular medicine, neurology, and genomics.

Harmanpreet Kaur is a Senior at the Medical Academy for Science and Technology in Homestead, FL. She is interested in majoring in Biomedical Engineering and Applied Mathematics, and pursuing an MD-PhD program. Her research interests involve genomics, chemical engineering, and medicine.
Analysis and Comparison of Safe Organic Ant Repellents

Jinghan Li
Beijing Haidian Foreign Language Shiyan School, NO. 20 Xingshikou Road, Haidian District, Beijing, 100195, China; 13070180372@163.com

ABSTRACT: Household ant invasion can cause both family health damages and financial losses. Professionally-used insecticides are not recommended for household usage because they can reduce human and animal health. Moreover, families lack knowledge on safe ant repellents because studies on organic ant repellents often showed inconsistent results. To provide concrete instruction on choosing safe organic ant repellents, this study investigated the topic in terms of the biochemistry of natural ant repellents that have been considered safe and effective, and provided a comparison among the repellents using a Y-tube olfactometer to determine the effectiveness of the ant repellents against two ant species. Spearmint, mint oil, and Chinese essential balm are shown to be effective ant repellents by using the plant alone or by mixing it with honey working as a bait. Although the choice of ants in the experiment does not represent all household-invasive ants, this study is still worth referencing for families plagued by ant invasions and for further studies on household ant control.

KEYWORDS: Animal Sciences; Other; Camponotus; Y-tube Olfactometer.

Introduction

Ant invasion, a notorious behavior of ant colonies of some species, including carpenter ants (Camponotus spp.), acrobat ants (Crematogaster spp.), and red imported fire ants (Solenopsis invicta), aiming for establishing new colonies or finding food and water sources, has been a common trouble in metropolitan and rural households.³ Household ants spoil fresh food, transmit diseases, sting people (which can be life-threatening for people with anaphylactic reactions), and even damage electrical currents as some species are attracted to electrical fields.²³ They are also a great nuisance as it is unpleasant for most people to see masses of ants marching in their home.³ These risks brought by invading ants result in increased health concerns and economic losses for the household. Professionally-used insecticides, albeit being the most effective agents in controlling ants, are not applicable to households, because they often negatively affect the health conditions of humans with their residual toxicity.⁴-⁸

It is known that the volatile odors of plants are effective in repelling ants, as the ants are naturally selected to be sensitive and respond to the chemical compounds in many plants.⁹ Many herbs are common therapeutic agents and ingredients used in cuisines and thus can be easily obtained.¹⁰-¹² Even though the organic materials (which were selected in the experiment, see details in methodology) may induce allergic reactions and even be toxic when used inappropriately for certain individuals, they are safe for the majority without allergies to these products and can still be applied by following the correct instructions. Besides, there was more than one product proven effective by the following experiment (view details in results and discussion section), so alternative choices are given for people who are allergic to one particular material.¹³-¹⁷ Therefore, these natural products are potentially safe organic choices for households to repel ants. However, previous studies mainly focused on the efficacy of one organic product in particular on repelling ants, but not on the effects of multiple potential ant-repelling products in comparison. Currently, organic options for household ant control are often related to the use of pepper, garlic, and vinegar. But these means have obvious drawbacks: pepper has irritating properties and can be dangerous for people with specific allergies; garlic has an unpleasant odor that lasts for hours; and vinegar may gradually erode objects if its application is not cleaned completely.¹⁸-²⁰ Therefore, finding safe organic ant repellent alternatives is important for controlling household ant invasions. This study tested the effects of nine organic insect-deterrents on repelling ants using Y-tube olfactometer and investigated the biochemical mechanisms of the plant-based ant repellents.

Materials and Methods

Ants:

Two species of carpenter ants, Japan carpenter ants (Camponotus japonicus) and sugar ants (Camponotus nicobarensis), were ordered as two separate colonies from a pet shop. These two species were chosen for two reasons. First, they are easily obtained and maintained, requiring limited space and efforts to raise as colonies, indoors. Second, these species are also considered household pests in tropical regions. Although household ants in non-tropical regions belong to different genera, the use of two species of Camponotus in the current study is still able to provide a reference regarding the effects of repellents on household ants.²¹-²³

Y-tube olfactometer:

A Y-tube olfactometer with a base length of 5 cm, arm length of 5 cm, and internal diameter of 5 mm was used. This Y-shaped device enables the volatile odor of tested material at each terminal of the arm to disperse to the fork junction of the tube.²⁴
Repellents:

According to previous studies on the efficacy of plant-based insect deterrents, the following easily-obtained materials were chosen to test their repelling effects on two ant species: citronella (Cymbopogon nardus), cinnamon (Cinnamomum cassia), spearmint (Mentha spicata), patchouli (Pogostemon cablin), clove (Syzygium aromaticum), ginger (Zingiber officinale), limonene, Chinese essential balm, and mint oil.²⁵-³¹ Citronella, cinnamon, patchouli, clove, and ginger were purchased from local grocers. Mint oil, Chinese essential balm, and limonene were purchased from household supply stores (the type of Chinese essential balm used was in liquid form with green coloration). Garlic (Allium sativum) and black pepper (Piper nigrum) were not examined here because of their unpleasant odor or irritating properties despite of their ant-repelling potentials and accessibility. Each repellent was placed at the terminal of one arm of the Y-tube olfactometer. The solid repellents (i.e., citronella, cinnamon, spearmint, patchouli, clove, and ginger) were ground into a powder first, and then 10 mg of each was applied in each trial (the powder was replaced after each trial). For the liquid repellents (i.e., limonene, Chinese essential balm, and mint oil), the volume of 50 μL was applied in each trial.

Stimulations for ant’s movement:

Two stimulants, honey and strong light, were applied in the experimental arena to ensure the movement of tested ants. Honey, with the volume of 50 μL, was placed at the terminal of each arm to attract the ants at the base to move toward one of the two arms. A torch illuminating the base of the Y-tube also served to propel the ant to make a choice rather than to remain still at the base of the tube. The light stimulation method was modified based on the observation that Camponotus spp. tended to stay in the shade rather than under the sun. The room temperature throughout the experiment was held constant at 26 ºC, because Camponotus spp. are most sensitive and responsive to the odor of food (honey) under this temperature.³²,³³

Experimental design:

One ant was placed at the base of the Y-tube and allowed to choose between the arms with and without the repellent (Figure 1). Each ant was tested only once and replaced with a new individual without exposure to the experimental arena because ants’ sensory system is able to remember recent odors and corresponding actions.³⁴ If the ant stayed in the arm without the repellent for 10 seconds, the tested repellent was regarded as effective, and vice versa. If the ant stayed still at the base of the arm for 2 minutes regardless of the stimulations of strong light and honey, the repellent was regarded as strongly volatile and powerful that even a tiny amount of its odorous particles could spread to the base of the tubes and discouraged the ant from moving forward. The effectiveness of each repellent was determined by the number of ants that made a particular choice: the more “no choice” and “repelled” occurred, the stronger the repellent was. Since the arms of the Y-tube are identical, the control group would only have to manifest that the ants would leave the base as the experimental design intends. However, the effect of the two propulsions for leaving the base, light and honey, has been well proven by prior research and experiments. Moreover, no other factors besides two propulsions and the repellent influence the choice of the ants. Therefore, a test with control group was not carried out. Still, it is necessary to stress that removing the control is only to be done with sufficient confirmation for the fundamental principle that the control intends to prove.

Figure 1: Experimental arena.

In the preliminary experiment, six C. japonicus and six C. nicobarensis were tested individually against nine repellent materials. The experiment was conducted without the stimulation of honey but with the stimulation of the light from the torch. The arena was also placed indoor with dim light to minimize the effect of vision (colors & shapes) on the ants’ choices.

In the first formal experimental phase, the two ant species were divided into four separate groups (each with six ants) for each repellent tested to determine if starvation would force ants to make choices: starving (no food or water provision for 24 hours) C. japonicus (abbreviated as SCJ), full (just fed until the ants refused to eat or drink) C. japonicus (FCJ), starving C. nicobarensis (SCN), and full C. nicobarensis (FCN). The formal experiment was conducted with stimulations of both honey and strong light.

In the second formal experimental phase, the experiment was repeated with only SCJ and SCN groups, because there was a significantly smaller number of “no choice” for the starving groups than the full groups in the first formal experiment (see details in the Results and Discussion section), suggesting that ants were willing to make the greatest effort on choosing one of the two arms when they are starved. This starvation pre-treatment could ensure that ants choosing to stay at the base were indicative of the strong volatility of effective repellents, thus, the individual ant staying at the base without a choice was regarded as strongly repelled ones.

In the third formal experimental phase, because an inconsistency of test results (see details in the Results and Discussion section) of the Chinese essential balm existed – the Chinese essential balm was a powerful repellent in the preliminary experiment but not effective in the two formal experiments – additional experiments were conducted using essential balm with 12 ants in each group (SCJ and SCN). The apparatus of adding the liquid repellent was switched from a graduated cylinder (error of 0.1 mL) to an injector (error of 0.02 mL) because it was hypothesized that the result
inconsistency was due to the inaccuracy of the apparatus used for adding the essential balm.

Results and Discussion
The preliminary experiment without stimulation of honey or starvation treatment:

As shown in Figure 1, among the nine plant materials tested, mint oil, spearmint, limonene, and Chinese essential balm (represented as CEB) were the most powerful repellents against both C. japonicus (represented as CJ) and C. nicobarensis (represented as CN), intuitively having more blue columns (signs of effectiveness) over red (signs of no effect) than other repellents did. However, there were some oddities regarding the data collected: given that “no choice” results from the repellent being very potent thus discouraging the ant from going even to the fork junction of the Y-tube, it is not logical for a repellent with a minor odor like cloves to have the same effect as the repellent with strong odor like Chinese essential balm does. Through further investigation, the reason for this “no choice” phenomenon could be explained by the light illumination: the torch created a strong halo at the edge of its circle-shaped illumination and gradually lost battery power during the experiment; so, the edge of illumination was the strongest, and the illumination was unstable (Figure 1). Therefore, when the illumination was provided by the torch, the ants were discouraged from leaving the base because the unpleasantly strong light got intense when they approached the fork junction. The data collected from the choice of C. nicobarensis also corresponded with the observation that they were not attracted to illuminated environments during the ant colony maintenance phase. Therefore, a new torch with even illumination and sufficient power was selected. Honey, in addition, was used to further propel the ants to leave the Y-tube base. Further experiments were not carried out in the dim environment anymore. Although the reason for conducting the preliminary experiment in a dim environment was to avoid the interference of vision (color and shape of the repellent) on the choice of ants, the reflections of the Y-tube illuminated the repellent at the end of the arm after the stimulating torch was applied. Therefore, eliminating the vision interference factor was not applicable anymore.33

The three formal experiments with stimulation of honey and starvation treatment:

After modifying the preliminary experimental method, results from the three formal experiments showed consistency in the effectiveness of spearmint and mint oil against both ant species with and without starvation treatment. Cinnamon, clove, and limonene also exhibited repellency against C. nicobarensis but not against C. japonicus, suggesting the differences in repellent sensitivity between ant species (see Figure 2 and Figure 3).

Chinese essential balm, which was a powerful ant repellent in the preliminary experiment (see Figure 1), was not consistently effective in the first or two formal experiment (see Figure 2 and Figure 3). Yet all ants that were “not repelled” by the essential balm died after eating the honey mixed with the essential balm, indicating the toxicity of essential balm. Thus, the third formal experiment was performed again for the Chinese essential balm only (see details in Methods section). As shown in Figures 4 and 5, the result differed from the first and second formal experiments: all tested C. japonicus and C. nicobarensis individuals were repelled by essential balm or had no choice (regarded as strongly repelled by essential balm, see details in Methods section).

There is a reasonable interpretation for this oddity. It was observed during the experiment that when ants chose the optimal position in the Y-tube to stay, they often exhibited tasting behavior on the honey and sometimes on the repellent as well. This seems strange because the ants can determine the odor via airborne particles to make a choice rather than by ingestion. However, the tasting behavior may happen because of the lure of honey. If ants tasted the honey in the arm with repellent, the ants tended to leave the arm immediately. For
the essential oil, however, the ants were killed shortly after tasting. This explanation is somehow consistent with the Chinese essential balm’s active ingredients: eugenol with LC50 value of 0.012 mg/cm² against carpenter ants, menthol with a significant influence on the foraging trails of ant colonies, and other components like eucalyptol and camphor, suggesting the effectiveness of essential balm in repelling and killing carpenter ants.⁶⁻³⁹ Indeed, facts mentioned above cannot be the definitive proof for my hypothesis as these testing on repellence was done via disturbing the foraging trail of a colony of ants, rather than single ants’ choice in the present experiment.⁴⁰ The exact repellency of the Chinese essential balm as a whole against carpenter ants was therefore not obtained in the current study, meaning that before considering how to use the essential balm, we must first consider whether a colony of ants acts in the same way in front of essential balm or its bait as a single ant in the experiment.

If it is confirmed that a colony acts as single ants did in the experiment, the next step is to determine the lethal dose of Chinese essential balm. The Chinese essential balm should be powerful enough to kill the ants when they taste the mixture, while they should still allow the ants to approach. The current study showed that when the percent composition of essential balm in the mixture was 50% (50 μL of honey and 50 μL of essential balm), the ants were not killed immediately after tasting; when the percent concentration of essential balm in the mixture reached at 75%, the ants were killed instantly after tasting. However, since the lethal dose of essential balm and the effectiveness of essential balm on other ant pest species like Coptotermes formosanus are unclear, further investigations need to be done. In addition, essential balm evaporates quickly, so its persistence of toxicity used as ant bait also needs to be determined in future.

Advantages of Chinese essential balm and spearmint include their security for household application and effective repellence upon other insect pests, such as mosquitoes and flies, due to the menthol component. Also, essential balm can even relieve the discomfort from mosquitoes’ bites.³¹⁻⁴⁴ However, as repellents, if not as bait, the essential balm and spearmint require further investigations as well: aside from testing whether the colony would be repelled by the repellents, ways of displaying the repellent require more study. The Chinese essential balm is highly volatile, so a thin, deep container would be required to achieve persistence of effect. Spearmint plant can be maintained indoor with sufficient light to provide repellence against ant pests: leaves at the bottom of the plant are often torn off to guarantee the growth of the plant, these leaves can be used to provide a repellent effect. Future studies also need to examine the repellent persistency of ground mint. If ground fresh mint fails to provide persistent repellence, other means, such as cutting the leaves in large pieces or slightly squashing the leaves, need to be examined.

■ Conclusion

The primary focus of future studies should be focused on essential balm and spearmint, as these two have the greatest prospect in application. Mint oil, another product that was proven effective in the study, is considered the essence extraction of mint plants, so the repellency of spearmint can be referred to mint oil, except that mint oil would require a container because of its volatility. While cinnamon, clove, and limonene were effective against only C. nicobarensis, they may not be the optimal choice for ant repellents because of limitation. At last, the verification of the repellent effect of the Chinese essential balm against a colony (rather than individual ants), amount of essential balm adequate for a bait, the effectiveness of spearmint plant alone as the repellent against a colony, and the universality of these two repellant materials, are worthy of future investigation.

■ Acknowledgements

I would like to thank my academic mentor Dr. Yao Xu for her guidance on methodology development, background knowledge about scientific research, and manuscript preparation.

■ References

Journal of the Canadian Dental Association 2008, 74(5).
33. Lopatina, E. B.; and V. E. Kipyratkov. The Influence of Temperature on Brood Development in the Incipient Colonies of the Ants Camponotus herculaneus (L.) and Camponotus xerxes Forel (Hymenoptera, Formicidae). Proceedings of the Colloquia on Social Insects 1993, 2
42. Abagli, A. Z.; Alavo, T. B. C. Essential Oil from Bush Mint, Hyptis Suaveolens, is as Effective as DEET for Personal Protection against Mosquito Bites. The Open Entomology Journal 2011, 5(1).
ABSTRACT: Starting at the end of January 2020, the global outbreak of COVID-19 has changed the lifestyles of many people due to its rapid person-to-person transmission. To reduce the spread, countries have taken numerous approaches, among which testing, vaccination, sanitization, lockdowns, and construction of quarantine centres are commonly implemented. Nonetheless, all these measures are expensive in terms of resources and have an effect on controlling the spread of the virus, quality of life, resource consumption and economic growth. In this essay, a novel intelligent method based on reinforcement learning (RL) is developed to provide recommendations on the optimal level of control measures, such as travel restrictions and lockdown policies, with the purpose to prohibit the further spread of the pandemic. By specifying continuous action space and defining customized reward functions, we provide a new learning framework to study such types of policy control problems, along with discussions on their theoretical motivations. Experiments practiced with actual COVID-19 data demonstrates that the suggested deep reinforcement learning algorithm based on the Deep Deterministic Policy Gradient (DDPG) model performs better than alternative RL algorithms as well as actual control measures.

KEYWORDS: Reinforcement Learning; COVID-19 optimal control policy; Continuous action space; Deep deterministic policy gradient.

Introduction

Human civilization has witnessed several pandemics in the past, including plagues, leprosy, smallpox, tuberculosis, AIDS, cholera, and malaria, see Rajaei A et al (2019). The historic records of pandemics suggest an implicit pattern that the frequency of disease outbreak increases as communication between civilizations grows, and it can therefore be reasoned that at this point of time when globalization is accelerating at an unprecedented rate, such contingencies are more likely to occur in the foreseeable future, as suggested in a WHO report in 2016. Thus, it is quite imperative to review the lessons learned out of our experiences with the current COVID-19 global pandemic in order to build a resilient society with people prepared to combat the social, health, and economic impacts of pandemics. Preparedness is a key factor in mitigating pandemics. It encompasses inculcating awareness about the outbreak and fostering response strategies to avoid loss of life and socioeconomic havoc. While the emergence of a harmful microorganism with pandemic potential may be unavoidable, pandemics can be prevented. Preparedness includes technological readiness to identify pathogen identity, fostering drug discovery, and developing reliable theoretical models for prediction, analysis, and control of pandemics.

The COVID-19 pandemic has led to numerous difficult situations. For now, there hasn’t been any 100% reliable immune vaccine to protect humans from infection. Some popular approaches, which are designed to control the spread of the infection in the current circumstances, are to enhance the frequency of testing as well as to broaden the coverage of vaccination in society. However, having all individuals tested and vaccinated is a time- and labour-consuming task. Besides being constrained by the available resources, governments also struggle in balancing the trade-offs between disease control and economic recovery. Another common measure is stricter sanitization in crowded locations, so as to disinfect public areas and thus control the spread of infection. However, this is also a resource expensive measure. It would be impossible to put in place sanitization at every corner and street of the country. On the other hand, lockdown approaches and quarantine plans are measures that rely comparatively less on extra consumption of medical, physical, or human resources. However, under current economic situations, it is unlikely to suspend various ongoing economic activities of the country and rely on the government to support and take care of its people, providing necessities for the mildly infectious people and keep available mission-critical ventilators for patients in severe conditions. Obviously, it is not possible to allow open and mask-free social gatherings, as the chance of spread is very high. There may exist other techniques that are not currently explored or discovered. Despite the ongoing techniques, including testing, sanitization and social distancing/lockdown/quarantine, there still needs to be an optimal level for each parameter in order to protect people from the virus and support routine activities while minimizing impacts caused on quality of life and economy. This paper focuses on tackling this problem using a quantitative, model-based approach. Specifically, we will build a reinforcement learning agent to formulate an
optimal policy, which will recommend the appropriate degree of regional lockdown and travel restriction respectively.

**Literature Review**

In this section, we will discuss the use of reinforcement learning algorithms in recommending COVID-19 related policies and introduce related research on this front. Reinforcement learning is a recent and popular research field, with the potential to make a resounding and laying impact on mankind’s history. It is a robust framework for optimizing specific tasks autonomously and has attracted great attention in both research and industry. The creation of intelligent machines using reinforcement learning will likely drive the understanding of human intelligence to places we have never been before. Intelligence, under this circumstance, is addressed as the ability to conform and apply knowledge acquired through past experiences. Arguably, if we can understand how to reach the optimal decisions for each and every problem, we likely understand the algorithm that recommends the optimal decision outcomes. An optimal decision may not necessarily be equal to maximizing immediate returns at hand; it may well represent the capability of trading not only immediate rewards for long-term goals, but also possibly a good, certain future. A good example would be joining a startup where a lot of factors are still uncertain instead of staying in a stable position. Goals that require a much longer time to materialize and have uncertain long-term value are usually the most difficult to achieve. Different from supervised and unsupervised learning, reinforcement learning is self-renewative by continuously interacting with the environment and resetting its actions based on reward signals. It has been confirmed that it can deliver a record high performance in games such as GO or DOTA, and it can be of more practical significance when combined with deep neural networks.

It is a nontrivial task to determine the proper level of policy measure for each country, largely due to the different demographic, social and economic structure in different countries and areas. Due to the nature of the disease, it is challenging to trace the source of an infection case, thus adding more difficulty and uncertainty in policymaking. Therefore, the government is faced with partial knowledge and incomplete picture about the status of the disease, when trying to identify the optimal restrictive policy. Such knowledge will only grow through a long-term study of the virus and its clinical characteristics. In addition, there is often a lag between infection date and reporting date, depending on the recording process in each region. Insufficient medical equipment, testing kits and vaccination provisioning have led to almost 3 million deaths globally as of mid-April 2021, which is also likely conservative due to disparate, and even purposefully underestimated reporting and information sharing. This pandemic has led to disruptions in almost all aspects of lives and businesses in society.³

Public health measures, including regional lockdowns and travel restrictions, have been analysed in the past to study their impact on controlling the spread of infectious disease. A typical research direction is the simulation of the spread of epidemics, which is a flexible framework to adjust to different disease structures and local intervention policies. This study has shown utility in simulating the spread of diseases and estimating the resulted impact. As for further targeted study on specific restrictive measure, Chizhov et al (2019)⁶ reviewed the effect of travel restrictions on the spread of COVID-19 and found that it is an effective measure in reducing the number of imported cases to other countries up to 77%, while only playing a limited role in controlling domestic spread. On this front, pre-emptive lockdown measures seem more promising in terms of slowing down and even reducing local transmission, as shown by Tian H et al (2020).⁶ When only a limited proportion of the total population is infected, the lockdown measures could play an important role in reducing the number of death cases, as shown by the simulations by Khadilkar H et al (2020).³ In their work, a reinforcement learning agent was constructed to learn an optimal policy that recommends quantitative lockdown measures, in consideration of health and economic impact. These interventions have proven to be effective at an aggregate level, but a more customized approach is still needed for individual countries and areas, with their unique demographic and socioeconomic characteristics being significant consequences of COVID-19 pandemic.

The complex dynamics among multiple factors caused by COVID-19 requires a robust and data-driven approach in order to formulate optimal preventive measures. To this end, Kwak GH et al (2021)⁸ proposed the use of deep learning approaches to recommend global public health strategies for COVID-19 pandemic. In their work, Duelling Double Deep Q-Network (D3QN) was used to discover optimal lockdown and travel restriction policies for individual countries and regions. However, the recommended actions are in a discrete space with a fixed number of values, which are still quite limited in applying to all countries and regions. A continuous action space is thus recommended, so as to allow for a more flexible and targeted policy recommendation. In our paper, we extend the work of Kwak GH et al (2021)⁸ and design a continuous action space for the reinforcement learning setting, where the agent is allowed to give an optimal action for lockdown and travel restriction, both varying within a fixed range and taking a numeric value, instead of a limited discrete space. In addition, we also deploy a more systematically advanced reinforcement learning algorithm named Deep Deterministic Policy Gradient (DDPG), whose superior performance is shown via experiments.

It is also worth noting that other research has studied COVID-19 pandemic using deep reinforcement learning from other angles. Regina P et al (2021)¹¹ proposed incorporating healthcare system parameters besides disease related characteristics to train a control model based on its transmission dynamics. Mahdavi M et al (2021)¹² explored COVID-19 related mortality risks using common machine learning algorithms such as Support Vector Machine and found that a small set of non-invasive features are predictive, indicating the potential use of these features in predicting intervention subject and location. Bednarski BP et al (2021)¹³...
analysed the use of reinforcement learning and deep learning models in optimizing the redistribution of medical equipment, thus instructing humanity to become better prepared for similar public health crises such as COVID-19. To the best of our concerns, none of the existing literature has studied the use of reinforcement learning in recommending optimal policies on lockdown and travel restriction with a continuous action space. Our major contribution is summarized as follows.

- We raise a new reinforcement learning framework based on continuous state and action representation space. This is more flexible than the discrete space used in Kwak GH et al (2021), since a continuous action space allows for more flexible and individualized policy recommendation on regional lockdown and travel restrictions.
- We introduce a new formulation of the optimal policy recommendation problem, based on a customized definition of the learning environment, including components on action and state space, as well as reward function. As we will show later, the customized environment is a flexible design which can be further tailored for more targeted and cost-sensitive learning.
- We proposed the use of the advanced DDPG algorithm in training of reinforcement learning agent and corroborated its superior performance by running experiments and comparing the results with multiple baselines. In addition, we also provide proof on the policy gradient theorem, a key component in DDPG architecture.

## Methods

### Basics of reinforcement learning:

Deep reinforcement learning is a genre of algorithms within the field of machine learning, aiming to solve artificial intelligence problems. It works by creating computer programs, or agents, to solve problems that normally demand intelligence. Compared with other types of machine learning algorithms, its uniqueness lies in its learning framework. Specifically, it learns through trial and error by interacting with the external environment and collecting feedback, including state and reward signals. This means that there is no labelled data, or an explicit correct answer to work with, which is different from supervised learning. The feedback from the environment could be simultaneously sequential, evaluative, and sampled from an original distribution. Powered by deep neural networks, an agent could be trained to approximate the true reward distribution using non-linear function approximation, with the ultimate goal of maximizing the long-term returns.

The goal of reinforcement learning is to train a decision-making agent that could achieve its target (maximizing cumulative rewards) despite existing uncertainties in its environment. At each time stamp $t$, an agent has a combination of action $a_t$ and state $s_t$ along with reward $r_t$ for each step. In a typical interaction process, the action $a_t$ is sent to the environment, which moves to the next state and provides reward $r_t$ to the agent. In this way, a reinforcement learning agent tries to maximize the cumulative rewards received after a certain policy, an entire series of actions, is performed. One
typical example that involves evaluative feedback in deep reinforcement learning is a Pong-playing agent; the agent lives in a Pong emulator and plays the game many times, learns by taking different actions and observing the corresponding effects. The agent will successfully play Pong with superior performance after training for many iterations.

Deep reinforcement learning focuses on creating computer programs, or agents, that can tackle complex and sequential decision-making problems under uncertainties. Notably this is a topic of interest in many fields, such as control theory and operations research. The trained agent will act as a decision maker, which will provide suggestions on what actions to take given a specific situation, or state from the environment. The environment is everything apart from the agent, over which the agent has no control. It is represented by a set of variables that characterize and make up the environment. The set of variables and all the different values that they can take jointly form the state space. One state is an instantiation, or a sample, from the whole state space. It is obvious that the agent may not necessarily have access to the full state of the environment. For states that are accessible, i.e., observations, are used to train the agent. This is also referred to as partially observable state space. At each state, the agent will have a set of actions to choose from, by mapping an input state to an output action. The agent influences the environment by taking a specific action. As a result, the environment may change its feedback, including state and reward, in response to the agent’s action. The transition from one state to another, based on a particular action from the agent, is referred to as the transition function, and is usually unknown to the agent. Depending on the type of algorithm, the transition function may be learned to better approximate the dynamics of the environment. Besides, mapping a specific set of state and action to a reward is referred to as the reward function. The set of transition and reward functions jointly form the model of the environment, which again is not available in most learning problems. The figure below (Figure 1) depicts the typical interaction cycle between the agent and the environment.

![Image](https://example.com/figure1.png)

**Figure 1**: Reinforcement learning interaction cycle.

The reward signals from the environment can be simultaneously sequential, evaluative, and sampled. Therefore, to obtain higher rewards, the agent needs to perform long-term thinking, balance exploitation and exploration, and finally generalize well the future scenarios. It typically follows a three-step process: first the agent interacts with the environment by taking an action, obtaining the next state and corresponding reward, then the agent evaluates...
agent improves its actions by changing its internal algorithmic representations of the world and all actions, expecting higher rewards in the next interaction cycle. The mapping from state to action is called policy, the mapping from state to a scalar value is called state value function, and the mapping function from a state action pair to a scalar value is called action value. Depending on specific requirements, the agent may choose to learn either an optimal policy or an optimal value. As we will introduce in the following sections, the type of algorithm we used in our approach employs both policy- and value-based methods, which has been shown to deliver superior performance compared with policy or value-based methods alone. We will also specify what constitutes the environment and the agent in our COVID-19 simulation in later sections.

In a nutshell, the optimal policy is automatically learned based on our proposed algorithm, which works as a function of environmental state, in particular disease parameters, such as number of deaths and infections, along with population characteristics such as density.

To facilitate our discussions in the following, we provide a table containing the mathematical notations used in this paper (Table 1).

Table 1: Mathematical notations used in this paper.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>(s \in S)</td>
<td>States</td>
</tr>
<tr>
<td>(a \in A)</td>
<td>Actions</td>
</tr>
<tr>
<td>(r \in \mathbb{R})</td>
<td>Rewards</td>
</tr>
<tr>
<td>(G_t)</td>
<td>Long-term returns, as a sum of immediate and future discounted rewards; (G_t = \sum_{i=t}^{\infty} \gamma^i R_{t+i})</td>
</tr>
<tr>
<td>(P(s', r(a, s)))</td>
<td>Transition probability from the current state (s) and action (a) to next state (s')</td>
</tr>
<tr>
<td>(\pi(a</td>
<td>s))</td>
</tr>
<tr>
<td>(V(s))</td>
<td>State value function measuring the expected long-term returns or value of being in state (s)</td>
</tr>
<tr>
<td>(V^\pi(s))</td>
<td>The value of state (s) following policy (\pi); where (V^\pi(s) = E_{\pi}[\sum_{t=0}^{\infty} \gamma^t R_t</td>
</tr>
<tr>
<td>(Q(s, a))</td>
<td>Action value function measuring the expected long-term returns or value of being in state (s) and action (a)</td>
</tr>
<tr>
<td>(Q^\pi(s, a))</td>
<td>The value of action (a) in state (s) following policy (\pi); where (Q^\pi(s, a) = E_{\pi}[\sum_{t=0}^{\infty} \gamma^t R_t</td>
</tr>
<tr>
<td>(A^\pi(s, a))</td>
<td>Advantage function (A(s, a) = Q(s, a) - V(s)), which has lower variance by extracting the baseline state value from the action value</td>
</tr>
</tbody>
</table>

In the following section, we first review the integral components of DDPG, including Deep Q Networks and Policy Gradient methods, followed by an introduction of the algorithm used in our experiment.

**Deep Q Networks:**

Since reinforcement learning problems can have both continuous state and action spaces, which can be either high dimensional with discrete values or low dimensional with continuous values, it is important to use function approximation that generalizes unseen states or actions, making the algorithm more efficient. In deep Q networks, as proposed in the seminal work by Minh et al. (2015)\(^{14}\), was used to generalize past experiences to new situations with superior performance, by approximating action value function using \(Q(s, a, w)\), where \(w\) denotes the weight of a neural network. Following a states-in-values-out architecture, the network could even yield high-performance implementation that outputs the value for all actions at once, given a specific state \(s\).

The objective function given by the action value network is as follows.

\[
L(w) = E_{s,a}[Q(s, a; w) - Q(s, a; w')^2]
\]

Where the objective is to minimize the loss with respect to the optimal action value function \(q_\pi(s, a)\) for the current state \(s\) and action \(a\), using the current estimate \(Q(s, a, w)\) parameterized by \(w\). The optimal action value \(q_\pi(s, a)\) is defined as follows.

\[
q_\pi(s, a) = \max_a_E[Q(s, a; w') - Q(s, a; w)]
\]

This essentially means the optimal action value function is the policy that renders the maximum expected returns from each and every action in each and every state. In the meantime, \(q_\pi(s, a)\) itself is not available in advance, it is often estimated using another target network, serving as a proxy target value from the target network for current estimation. In deep Q networks, this is represented using an online Q-learning target, defined as follows.

\[
y = r + \gamma \max_a Q(s_{t+1}, a' | s_{t+1})
\]

Which is a combination of the actual experienced rewards and the estimated action values of the next state. Plugging in \(y\) from the target networks gives the following.

\[
L(w) = E_{s,a,r,s'}[(r + \gamma \max_a Q(s', a'; w') - Q(s, a; w))^2]
\]

Where \(r\) represents the parameters of the online Q-learning network, and \(w\) denotes the parameters for the target network, which is usually updated every few iterations. By differentiating through this equation to update the network weight using gradient descent algorithm, we have:

\[
\nabla_{w}L(w) = E_{s,a,r,s'}[(r + \gamma \max_a Q(s', a'; w') - Q(s, a; w))\nabla_{w}Q(s, a; w)]
\]

It is worth noting that the gradient does not involve the target, thus it only goes through the predicted value, namely \(Q(s, a, w)\).

**Policy Networks:**

Since reinforcement learning problems can have both continuous state and action spaces, which can be either high dimensional with discrete values or low dimensional with continuous values

\[
J(\theta) = \sum_a d(\theta, a | s) - \sum_a d(\theta) \sum_a d(a | s, \theta)
\]

Where \(d(\cdot)\) denotes the on-policy stationary state distribution of the Markov chain for \(\pi_\theta\) under policy \(\pi\). In other words, \(d(\cdot) = \lim_{t \to \infty} P(s_t = s | s_0, \pi_\theta)\) is the final stationary probability that \(s_t = s\) when starting from initial state \(s_0\) and following policy \(\pi_\theta\) for a total \(t\) steps. The parameters \(\theta\) of the policy network can then be sequentially optimized using gradient ascent algorithm, based on the gradient \(\nabla_{\theta} J(\theta)\).

It is challenging to compute the gradient \(\nabla_{\theta} J(\theta)\) due to its dependence on both action selection \(\pi_\theta(a | s)\) and the stationary distribution \(d(\cdot)\). Based on the work of Sutton & Barto (2017),\(^{15}\) the derivative \(\nabla_{\theta} J(\theta)\) can be transformed to a form that does not involve the derivative of state distribution \(d(\cdot)\). We build on the work of Sutton & Barto (2017)\(^{15}\) and provide

DOI: 10.36838/v4i3.11
the following theorem to characterize the policy gradient used in our algorithm.

**Theorem 1:**
The gradient of $J(\theta)$ does not depend on $d_\pi$ and only depends on $\pi_\theta(a|s)$ through the following:

$$V_\pi(\theta) = \sum_{s,a} \nabla_\theta J_\pi(\theta) = \sum_{s,a} \nabla_\theta \mathbb{E}_{P_\pi(s'|s,a)}[\psi]\mathbb{E}_{P_\pi(s'|s,a)}[\phi].$$

**Proof of Theorem 1:**

We first analyse the derivative of state value function $V_\pi(s)$.

$$V_\pi(\theta) = \sum_{s,a} \nabla_\theta J_\pi(\theta).$$

From the previous objective function $J_\pi(\theta)$, we have the following recursive representation of the visitation probability:

$$p_\pi(s'|s,a) = \sum_{s'} p(s'|s,a).$$

In this stage, our goal is to continue transitioning from state $s_k$ to state $x$ at step $k$, following the policy $\pi_\theta$. When $k=0$, we have $P_{\pi}(s_k=0)=1$. When $k=1$, the transition probability is calculated by summing up all possible future transitions, weighted by possible action probabilities:

$$P_\pi(s \rightarrow s', k = 1) = \sum_{s',a} P_\pi(a|s)P(s'|s,a).$$

The above derivation essentially removes the dependence of the derivative on the action value, namely $\nabla_\theta J_\pi(\theta)$.

Algorithm 1: Deterministic Policy Gradient (DDPG)

1. Randomly initialize critic network $Q(s,a)$ and actor network $\pi(s,\theta)$ with parameters $w$ and $\theta$.
2. Initialize target networks $Q'_\pi(s,a)$ and $\pi'_\theta(s,\theta)$ with parameters $w$ and $\theta$.
3. Sample a random mini-batch of $N$ transitions $(s_t,a_t,r_t,s'_t)$ from $\mathbf{R}$.
4. Compute the gradient in the parameter $\theta$ for the target network $\pi'_\theta$.
5. Update the target network $\pi'_\theta$.
6. Update the parameter $\theta$ using gradient $

**Experiments:**

To verify the effectiveness of our proposed model, we conducted experiments using actual COVID-19 related disease data from 186 countries and regions around the world, in the period between 7 Jan 2020 and 22 Jan 2020. The data
consists of disease related features such as the number of confirmed, recovered, deceased cases per day, as well as region specific features such as GDP, population, and life expectancy, etc. Additional features are also derived, including the relative ratio and cumulative statistics of these features. Since the training environment for this problem is not provided in any of the existing open-source libraries, we created our own customized environment, with a reasonably defined reward function to support optimal decision making of the agent. Our framework, along with its multiple parameters for reward definition, is designed to be flexible enough to support future extensions. In the following sections, we will introduce the details on reward function definition, following illustrations of the experiment results and discussions.

**Reward definition:**

A proper reward function should be defined in order to guide the agent in the search of the optimal policy, in our case controlling the real pandemic situation by tuning the intensity of lockdown and travel restriction on a daily basis. Intuitively, a rise in the number of deaths or infections should be discouraged, thus resulting in a negative reward under such circumstances. On the other hand, when some actions have succeeded in controlling the number of infection and death cases under a certain level, then they should be encouraged, and thus the corresponding rewards should be positive. The reward used in our environment consists of two components: death and recovery, each of which taking a different reward or penalty structure. The total reward per day is then derived by summing up both components. The following table illustrates the overall sign direction for the rewards definition logic, based on death severity. The same logic applies to recovery, which will take an opposite sign.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Explanation</th>
<th>Reward sign</th>
</tr>
</thead>
<tbody>
<tr>
<td>Getting worse</td>
<td>More death cases than yesterday without any actions taken</td>
<td>Negative</td>
</tr>
<tr>
<td>Stable</td>
<td>No additional death cases than yesterday due to certain actions taken</td>
<td>Positive</td>
</tr>
<tr>
<td>Getting better</td>
<td>Reduced number of deaths due to certain actions taken</td>
<td>Positive</td>
</tr>
</tbody>
</table>

Table 2: Example of designing the reward sign direction for death severity. The same logic applies to death and recovery, with the latter taking the opposite sign. Note that due to the contagious nature of the disease, the number of deaths will almost surely increase if no action is taken. On the other hand, if the situation gets better because of certain actions taken, then its effect may only start to reveal after a few days, and the reward will eventually be positive and thus encouraged by the agent.

Based on this overarching rule, we designed the reward function $r_t$ as follows. Note that customized weighting between these two components is possible, with the flexibility of incorporating additional perspectives such as number of infections.

$$
\begin{align*}
    r_t^{\text{death}} &= -c_0 & \text{if } \text{death}_t > \text{death}_{t-1} \text{ and } \text{death}_t \neq 0 \\
    &= c_0 & \text{if } \text{death}_t = \text{death}_{t-1} \text{ and } \text{death}_t \neq 0 \text{ and lockdown} > 0 \text{ and travel restriction} > 0 \\
    r_t^{\text{recovery}} &= c_0 & \text{if } \text{recovery}_t < \text{recovery}_{t-1} \text{ and } \text{recovery}_t \neq 0 \\
    &= -c_0 & \text{if } \text{recovery}_t = \text{recovery}_{t-1} \text{ and } \text{recovery}_t \neq 0
\end{align*}
$$

**Action space:**

In this paper, we propose to use continuous action space to determine the intensity level for lockdowns and travel restrictions. The proposed values for these two action outputs vary within a fixed range, thus this is a more flexible and customized recommendation engine compared with discrete action space. From an algorithmic perspective, since a regular deep Q-learning network is unable to deal with continuous action space due to the curse of dimensionality, the DDPG algorithm used in our model could, by design, handle the continuous output space very well.

**State space:**

Since disease and region related features are time series in nature, it is important to expose the trend of these metrics as part of the observations in a state. For example, if we only provide the number of infections for the current day, the agent will have no indication of whether such a metric is improving or worsening. By incorporating prior information from previous days, we allow the agent to infer serial dependence structure from the multivariate time series data, thus aiding the long-term decision on whether to ramp up or ease certain control measures. In our experiment, we provided past 10 days of feature values as a single state input. In other words, the agent is exposed to a snapshot of previous 10 days of feature input values per day, on a rolling basis.

**Results**

By training a reinforcement learning agent using actual data and corresponding simulations via experience replay, the algorithm will return a policy that provides theoretically reward-maximizing parameters on the control policies under given states. Yet, in order to test its validity, the reward outputs from the algorithm-based policy should then be compared with the real reward outputs under the same state. Once most theoretical rewards exceed their corresponding original ones, it can be concluded that our agent managed to discover an optimal policy for pandemic control.

We first calculated the rewards based on actual lockdown and travel restriction and plotted the distribution of rewards via a frequency plot shown below. Figure 2 suggests that most of the countries and regions are taking insufficient lockdown or travel restriction measures in the fight against COVID-19. However, there are some countries showing good performance in controlling the spread of the disease, which corresponds with our actual observations.

As a comparison, we first implemented the Actor Critic algorithm using Kronecker-Factored Trust Region (ACKTR) method, as proposed by Wu et al (2017). This serves as a baseline model for comparison with our DDPG based model. We trained the agent using ACKTR algorithm for a total of 5000 epochs, and then used the trained agent to interact with the environment and generate a list of rewards, which were then added up to form a final cumulative reward. We repeated this experiment 100 times, and then compared it with the actual observed rewards by taking the difference between
them. Figure 3 suggests that the agent-based recommendation system can generate policies that are better than existing ones for most of the time. The superior policies also surpass the existing policies by a huge margin in terms of the cumulative rewards.

**Figure 2:** Reward distribution based on actual lockdown and travel restriction policies. The “Frequency” axis stands for the number of policies that with a certain interval of reward. The figure suggests that most rewards are negative indicating the largely insufficient actions taken by most countries and regions.

**Figure 3:** Distribution of reward differences based on lockdown and travel restriction policies recommended by ACKTR model and those from actual control measures. The figure suggests that the agent-based recommendation system is able to generate policies that are better than existing ones for most of the time. The superior policies also surpass the existing policies by a huge margin in terms of the cumulative rewards.

We also plotted the distribution of cumulative rewards side by side. Figure 4 also suggests that the trained agent using ACKTR can better recommend optimal policies that maximize the long-term returns.

**Figure 4:** Side by side comparison of density plot on cumulative rewards. The “Data” axis stands for corresponding reward value calculated with each given policy. The figure suggests that the trained agent using ACKTR is able to better recommend optimal policies that maximize the long-term returns.

Now let us visit the experiment results using DDPG model. Under the same overall setting, we set the discounting factor to 0.99, maximum number of episodes to 100, maximum mean reward per 100 steps to 10000, and the learning rate for both policy and value networks to 0.0001. We set the replay buffer size to 10000, with a batch size of 32. For the soft update parameter $\tau$, we adjusted it to 0.005. Figure 5 contains the moving average rewards for both training and evaluation, where the evaluation performance quickly stabilizes to around 6000 in early episodes and slowly improves afterwards, suggesting that the DDPG algorithm is efficient and effective in capturing the long-term systematic patterns in the data.

**Figure 5:** Moving average rewards for both training and evaluation using DDPG algorithm. The evaluation performance quickly stabilizes to around 6000 in early episodes and slowly improves afterwards, suggesting that the DDPG algorithm is efficient and effective.

Besides, we also compared its empirical performance with actual lockdown and travel restriction measures. By taking the difference between the cumulative rewards using DDPG and that from actual control measures, we have the following Figure 6, indicating a uniformly positive reward difference and thus superior performance.

**Figure 6:** Distribution of reward differences based on lockdown and travel restriction policies recommended by DDPG model and those from actual control measures. The figure suggests that DDPG model is able to generate uniformly superior control policies compared with actual lockdown and travel restriction measures.

- **Discussion**

Coming up with an optimal control policy that balances the immediate evolution of the situation and long-term impact is
a challenging task. One needs to consider not only the current factors such as infection status, but also the pre-existing trends of this metric, so as to properly adjust and optimally determine the next control policy. By proposing an automated framework to learn from the simulated environment based on actual data, it is thus our hope that the trained agent serves as a decision support engine to assist policymaking of local authorities, in this case the level of lockdown and travel restriction. Plus, the recommended policies should be flexible enough to cater to individual characteristics of different countries and regions.

Normally, the level of lockdown and travel restriction in a region is graded in a discrete classification. For instance, level 0 lockdown could mean no control measures at all, and level 2 of lockdown could suggest an entire shutdown of transportation routine. Nonetheless, it is unlikely for the government to exactly borrow and copy the standards of lockdown classification from other countries or regions. One major contribution of our work is to specify the actual levels of lockdown to be expressed as continuous (in this case, a floating number that lies on [0, 2], similarly for travel restriction measures). With the support of the DDPG algorithm, such outputs could more appropriately cater to practical implementations.

In addition, defining a proper reward function is of great importance to the successful training of a reinforcement learning agent. The reward serves as a weak signal. Compared with supervised learning, instead of directly indicating the correct answer to the current learning problem at hand, a reward signal simply indicates the immediate feedback from the environment. Due to the lack of transition dynamics of the environment, simply maximizing the immediate rewards may lead to early and sub-optimal convergence. Therefore, the reward definition should reflect the attributes of the environment which the agent is expected to learn from and maximize, at the right time, via a long planning horizon. Inappropriately defined reward functions may confuse the learning algorithm when trying to identify the right direction to improve, or in some cases, mislead the agent into leading something not in the right direction. In our paper, the customized reward function is defined in a way that it encourages early control of the disease and penalizes bad or even worse situations due to the lack of proper control actions. We believe such a definition meets the overarching objective of disease control.

Nonetheless, it should be noted that in real implementations, both the control measures and the spread of the pandemic lead to economic loss in society, which is not taken into consideration as part of our reward function definition. The potential downstream impact is that the optimal policy suggested by our agent will be too costly to practice, as it is more beneficial to choose the safe and conservative side of control, as early as possible. Incorporating the economic impact, along with the delayed influence on economic growth as a whole, would be an interesting future research direction.

■ Conclusion

In this paper, we study the problem of formulating an optimal policy for regional control measures on the spread of COVID-19, and propose the use of DDPG algorithm, a powerful reinforcement learning model, to tackle this challenge. We analyse the structure of the optimal control problem by detailed discussions on the definition of three integral components of reinforcement learning: action space, state space, and reward function. To demonstrate the effect of the DDPG model, we provide theoretical motivations on its value and policy networks, with arguments on its superior performance compared with alternative models. In addition, we perform experiments using actual COVID-19 data and show that our proposed model provides better empirical results in terms of total cumulative rewards, when compared with both actual control measures and those from an alternatively trained agent. By this process, our agent shows better results in terms of controlling the spread of the pandemic. These results suggest that DDPG is a promising model for similar types of control problem.

■ Acknowledgements

I would like to thank Professor Wu Yu for guiding me through the research process and teaching me how to write a research paper.

■ References


Author

Kailiang Liu is currently studying in Shenzhen Middle School, and is inclined to apply with the direction of mathematics and computer science in future college study.
Invisible Labour: Comparison of Unpaid Work Done by Women and Men across Developed and Developing Countries

Neasha Mittal

Delhi Public School, Sector 12, R. K. Puram, New Delhi, 110022, Delhi, India; neashamittal@gmail.com

ABSTRACT: The research topic addresses the phenomenon of the gross discrepancy between the quantity of unpaid care work between men and women throughout developed and developing countries. Specifically, this study compares the percentage of a day men and women spend in unpaid work across developed and developing countries through the analysis of secondary data obtained from reputable sources. The results showed that women in both developed and developing countries did a higher proportion of unpaid work in all countries; the disparity between the two genders was significantly higher in most developing countries, with a few exceptions. The disproportionate burden borne by the women among the three groups is also underscored by the fact that the proportion of the day spent in unpaid labour is the same for both developed and developing countries. Additional exploration of related literature suggests that socioeconomic and cultural characteristics of the developing countries, which place women in a disadvantaged position in terms of the opportunity for self-development, further accounts for these differences. Various factors, like culture, were identified to account for the differences. As such, the following recommendations were made institution of gender-equitable policies, free/affordable compulsory education for all, creation of a model to generate monetary values for unpaid labour.

KEYWORDS: Economic empowerment; developed and developing countries; gender inequality; unpaid care work; unpaid female labour.

Introduction

Invisible Labour: Comparison of Unpaid Work Done by Women and Men across Developed and Developing Countries:

It is a well-known, but oft-neglected fact that women all around the world, be it in developed or developing countries, carry out a greater share of unpaid care work responsibilities than men every day.¹ Globally, women carry out 76% of the total amount of this work, which is over three times more than men do.¹ Within the context of this research paper, unpaid work is largely centred upon the following two broad categories:

1. Non-market work:
   This category includes the production of goods and services that are not sold in the market. Subsistence agriculture, i.e., the production of food for own consumption, is a good example.²

2. Unpaid care work:
   The unpaid care work, also known as household work, domestic labour, or family work, encompasses the provision of services for family and community members outside of the market.³ Caring for other people’s children, or other family dependents (whether the sick or elderly), often represents a form of informal reciprocity that is repaid in kind by providing care for their family members in times of need.²

This unpaid care work that is critical to everyday life lies at the heart of the productivity of a nation. The quantification of unpaid care work and non-market work for the purpose of national income accounting is a complex issue obscured by its very nature and subjectivity. Many of the unpaid workers are engaged in activities that are fundamentally indistinguishable in most respects from paid employment. The unpaid work can’t always be put strictly into the category of work without remuneration. However, if a paid service provider were to take up a portion of this unpaid work, then this would ensure the delivery of professional care.² Many labour force surveys consider ‘unpaid family workers’ as those who contribute to a family farm or enterprise without receiving direct payment, as an indirect contributor to the family income, essentially acknowledging that they are contributing to the economy.² However, the lack of direct compensation means that their work is difficult to quantify. Another factor that complicates the quantification of unpaid labour is the concern for the well-being of care recipients, which often affects the quality of the services rendered.² In the absence of logically fair quantification of unpaid work, the imbalance in unpaid work distribution goes unnoticed by policymakers.

Women’s unequal share of unpaid care work has long been recognised by various social organisations as a key dimension of gender inequality. The patriarchal society considers unpaid care work as a female’s responsibility.³ Policymakers are unable to recognise the existence of an unpaid labour force, which is not only taken for granted, but also frees up other individuals for paid work. For instance, housewives taking up unpaid work (caring for young children and the elderly) enables their husbands to do full-time work and receive full compensation without having to worry about home care responsibilities. However, this places women in the disadvantageous position of financial dependence, which clearly sets the foundation for gender disparity. Women across different regions, classes, and cultures spend a significant portion of their day on meeting the expectations of their household duties.⁴ More than just an
issue of economic inequality, this unequal division of unpaid labour essentially limits women from achieving their potential in various spheres of life. Even when women do paid work, the expectation that they fulfil their unpaid activities limits their options on the paid activities due to these constraints. Therefore, their performance of paid work over and above the unpaid work simply creates a double burden for women.

The way society and policymakers have been addressing this issue of disproportionate involvement of women in daily unpaid care work also has important implications for the achievement of gender equality. They can either expand the capabilities and choices of women and men, or confine women to traditional roles associated with femininity and motherhood. Essentially, the unequal distribution of unpaid care work between women and men infringes on women's rights and constitutes a stumbling block on their economic empowerment.

Furthermore, the level of development of the economy has definite implications in precipitating gender inequality, as it affects not only the duration, but also the distribution, of time between paid/unpaid work and the allocation of unpaid time among a variety of activities. Specifically, the lack of an adequate public sector infrastructure and the absence of social service delivery services can also play a role in the specific allocation of time among a variety of unpaid tasks. As a consequence, women in rural communities of developing countries have to devote more time than their counterparts in developed communities to perform basic domestic tasks, such as fetching water or engaging in subsistence production. The provision of amenities and services, such as water delivery to one's doorstep, child and elder care, as well as universal free access to health services, would reduce the amount of time needed for women in developing countries to perform these unpaid tasks, thus reducing such an imbalance.

Moreover, in the developed economies, paid job opportunities are higher for a wider segment of the population. In fact, women in developed countries comprise a high percentage of women in the workforce: Sweden and Denmark, 75%; the United States, 60%; and the United Kingdom, 59%. Conversely, women in developing countries make up only 31% of the formal labour force, such as 33% in Brazil and Chile, as well as 32% in Mexico. Nonetheless, between 65% to 90% of all part-time workers are women, due to their unpaid child care and home-based responsibilities. Further exacerbating the lack of inequity in the workplace situation is the concentration of employment opportunities in a narrow range of sectors (especially services, where access to jobs is easier, but wages are often lower and job security minimal).

Despite the above-mentioned differentiating elements between the conditions of the women living in developing and developed countries, both groups perform disproportionately more unpaid work than their male counterparts. Today, in mainstream economic policy discussions, the disparity between unpaid care work taken up by men and women has been identified as a major constraint to both economic growth and women's economic empowerment.

This research study sought to offer a global overview of the situation through a statistical comparison of the inequality in the distribution of unpaid labour between males and females in a selection of developed and developing countries. The adoption of this broad overview of countries using the paradigm of developmental status was deliberate to evaluate the impact of socio-economic and cultural factors, as well as their interaction, by situating it within the context of an exhaustive review of the literature.

Methods

Research Aim & Research Approach:
The research aim of this study was to compare the proportion of daily hours of total unpaid care work by women and men among the developed and developing countries, thus ascertaining the relative impact of gender and development status. The hypotheses that would be evaluated are presented below:

1a. Null Hypothesis 1:
There is no difference between the mean percentages of daily unpaid care work done by females and males.

1b. Alternative Hypothesis 1:
There is a difference between the mean percentages of daily unpaid care work done by females and males.

2a. Null Hypothesis 2:
There is no difference between the mean total percentages of daily unpaid care work done in developed and developing countries.

2b. Alternative Hypothesis 2:
There is a difference between the mean total percentages of daily unpaid care work done in developed and developing countries.

3a. Null Hypothesis 3:
There is no interactive effect between gender and development status on the mean percentage of daily unpaid care work.

3b. Alternative Hypothesis 3:
There is an interactive effect of gender and development status on the mean percentage of daily unpaid care work.

Data Collection

To accomplish the objectives, the data from various sources both at the micro and macro levels were utilized in this study. This study has collected the data on proportion of time spent on unpaid domestic and care work from the following sources:

1. Developed:
   - Australia and New Zealand (South Pacific); Hong Kong and Japan (East Asia); Sweden, Norway, Denmark, Finland (Scandinavia); UK, Austria, Belgium, France, Germany, Ireland, Switzerland, Estonia, Greece; Poland (Western, Northern, and Eastern Europe); US and Canada (North America).

2. Developing:
   - Mexico, Cuba, Costa Rica, Argentina, Brazil, Bolivia, Chile, Colombia (Central and South America); Russia, Turkey, Bulgaria, Armenia, Kazakhstan (Northern, Western, Central Asia); India, China, Bhutan, Thailand, Cambodia (East Asia); Cameroon and South Africa (Africa).
Data Analysis:

A two-way ANOVA analysis was also run to determine whether mean differences between the percentages of daily unpaid care work by gender and developmental status are statistically significant and whether there is an interactive effect between the two factors. Furthermore, the implications of the inequitable distribution of unpaid labour between males and females were used as a springboard for further exploration within the context of the literature.

Results and Discussion

In this section, the differences between the mean percentages of the day spent in unpaid labour of males and females of a selection of developed and developing countries are presented and analysed in detail (see the raw data in the Appendix).

A Two-Way ANOVA with Replication was run to compare the differences between the mean percentages of day spent in unpaid labour between females and males in a selection of developed and developing countries, as well as determine their statistical significance. The results are presented in Table 1. The main effect of gender yielded an F ratio of 176.49, \( p < .01 \), which was substantial. It indicates that the differences in the mean percentages of day spent in unpaid labour between males (\( M = 9.13\% \), \( SD = 2.7\% \)) and females (\( M = 15.88\% \), \( SD = 2.37\% \)) in the developed countries, as well as the males (\( M = 6.2\% \), \( SD = 3.0\% \)) and females (\( M = 18.6\% \), \( SD = 4.43\% \)) in the developing countries, is statistically significant. As noted in the literature, females work more than males in both developed and developing countries. The disparity in developing countries is almost double that of developed countries.

Table 1: Two-Way ANOVA with replication — Mean percentages of day spent in unpaid work among females and males in developed and developing countries.

<table>
<thead>
<tr>
<th>SUMMARY</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Developed</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Count</td>
<td>20</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>Sum</td>
<td>182.5</td>
<td>317.5</td>
<td>500</td>
</tr>
<tr>
<td>Average</td>
<td>9.125</td>
<td>15.875</td>
<td>12.5</td>
</tr>
<tr>
<td>Variance</td>
<td>7.29565789</td>
<td>5.62928547</td>
<td>17.9789744</td>
</tr>
<tr>
<td>SD</td>
<td>2.70</td>
<td>2.37</td>
<td>4.24</td>
</tr>
<tr>
<td>Developing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Count</td>
<td>20</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>Sum</td>
<td>124</td>
<td>372</td>
<td>496</td>
</tr>
<tr>
<td>Average</td>
<td>6.2</td>
<td>18.6</td>
<td>12.4</td>
</tr>
<tr>
<td>Variance</td>
<td>9.01052532</td>
<td>19.62106286</td>
<td>53.374359</td>
</tr>
<tr>
<td>SD</td>
<td>3.00</td>
<td>4.43</td>
<td>7.31</td>
</tr>
</tbody>
</table>

These findings confirm the prevalent literature that women all over the world spend a substantial amount of daily time on activities that are not typically recorded as ‘economic activities’, in addition to their part-time or full-time job. Statistically, women spend an average of 4.5 hours a day on unpaid work. That is more than double the amount of time men spend. Hence, female participation in labour markets tends to increase when the time-cost of unpaid care work is reduced, shared equally with men, and/or made more compatible with market work. For instance, a typical Ghanaian woman, works on average, 13 hours per day. However, while 40% of her day’s work is paid, most of her work is unpaid. Conversely, men spend their non-work time on leisure activities, such as playing sports, watching TV, and hanging out with friends. Consequently, reducing the gender imbalance in the time devoted to unpaid care work will have a positive impact on reducing gender inequality in every sphere of life.

As for the main effect of development status, it is not statistically significant, with an F ratio of .02, \( p = .89 \). This statistical result is unsurprising, given how close the mean ratings are for the two groupings: developed countries (\( M = 12.5\% \), \( SD = 4.24\% \)) versus developing (\( M = 12.4\% \), \( SD = 7.31\% \)). This indicates that the mean percentage of a day spent in unpaid labour is largely the same, whether one is living in a developed country or a developing country. This finding thus challenges the literature that points out that time spent in unpaid labour in developing countries exceeds that of developed countries due to the absence of proper infrastructure. Rather, the actual inequity in unpaid labour stems from the fact that the responsibility is largely imposed upon women, particularly those in developing countries.

Nonetheless, the interaction effect (though smaller in size than in the case of gender) is still significant, \( F \) ratio = 15.36, \( p < .01 \). The statistical significance of the interaction effect between gender and developmental highlights the need to acknowledge the differences in the mean percentages of day spent in unpaid labour of females living in developed and developing countries, as well as the corresponding figures of their male counterparts. First, the mean percentage of the day spent in unpaid work among women (\( M = 18.6\% \), \( SD = 4.43\% \)) exceeds that of their counterparts in the developed countries (\( M = 15.88\% \), \( SD = 2.37\% \)) by 2.72%. In the meantime, the males in the developed countries (\( M = 9.13\% \), \( SD = 2.7\% \)) had a higher mean percentage of the day spent in unpaid labour than their counterparts in the developing countries (\( M = 6.2\% \), \( SD = 3.00\% \)) by 2.93%. Given the fact that both developed countries and developing countries had similar percentages of time spent in unpaid labour, it is evident that women in developing countries bear the greatest brunt of the labour among all three groups, thus highlighting the interaction effect of both gender and developmental status in influencing their situation with regards to unpaid labour.

A close examination of the differences between the percentage of the day spent in unpaid labour between the genders of individual developed and developing countries also generates additional insights. Figure 1 shows the striking differences between the mean percentages of the day spent in unpaid labour of women and men per day, particularly among the developing countries.
countries. In particular, developing countries like India stand out with the ratio of unpaid work of women to men standing at 6.83:1 (women: 24.46%, men: 3.58%), with Turkey having a ratio of 4.50:1 (women: 21.17%, men: 4.71%) and Mexico with a ratio of 2.52:1 (women: 23%, men: 9.13%). In stark contrast, among the developed countries, particularly those in Europe, the disparity between the genders is slightly better, with the ratio of the unpaid work of women to men going as low as 1.5 times in Norway (women: 15.83%, men: 11.71%). Denmark has a ratio of 1.3:1 (women: 16.87%, men: 12.92%), while France has a ratio of 1.66:1 (women: 15.54%, men: 9.37%).

However, among some developed countries, the ratio of the unpaid work of women to men can still go as high as 5.56 times, as in the case of Japan (women devoting 15.75% of the day, men 2.83% of the day) and 3.66 times in Hong Kong (women devoting 11% of the day, men 3% of the day). This reality among some developed countries highlights the prevalence of gender inequality across the globe in irrespective of higher economic development.

Male vs Female Hours of Unpaid Labour: Developed and Developing Countries.

**Figure 1**: Mean percentages of the day spent in unpaid labour of males and females in developed and developing countries. Globally, women take up a higher proportion of the unpaid work; the disparity is less in developed countries.

The significant gender disparity among the developed countries of the East Asia and Pacific region in comparison to the other higher income countries of the Western region or Europe may be attributed to the cultural foundation and its development over the centuries. Although the role that Japanese society ascribes to women has changed over the course of time, reflecting the trends of respective eras, the current expectations of Japanese women are still founded on historical norms. Before the Meiji era at the end of the 19th century, women were expected to stay within the home, as opposed to their husbands, who were expected to seek out jobs; this served as the foundation of the government and Confucian values, as each home represented a building block for the culture.¹⁴,¹⁵ Despite the changes in consumer culture and the modernization of Japanese industry, more emphasis was put into the upholding of the patriarchy, particularly to the adherence of familial roles and structures, contributing to an intensification of gender roles.¹⁶-¹⁸

When discussing gender roles, Dalton's research reveals that there is still a significant trend towards traditional gender role expectations in Japanese society, such as women being homemakers or, even more importantly, mothers, despite procuring outside employment or even positions in the Japanese government.¹⁹ The women in Japan choose to contribute to the economy via paid work alongside their male counterparts, as well as conform to these extremely traditional gender roles; therefore, the role of a modern woman is difficult in Japan as women try to balance personal, traditional, and community values.²⁰-²²

Such cultural issues with the specific region, irrespective of development, may be keeping a larger section of the women population away from the full-time paid labour market. This aspect of society needs to be addressed closely by policy makers to redress gender inequality in these countries.

**Conclusion**

The data analysis of the mean percentages of unpaid work per day of males and females in 20 developing and 20 developed countries shows that gender is the chief determinant of the proportion of time spent in unpaid labour, regardless of the development status of the countries. Specifically, women spend a greater proportion of time in unpaid labour than their male counterparts. Therefore, the null hypothesis that there are no differences between the mean percentages of unpaid work per day between the genders in both developed and developing countries can be rejected. Women in developing countries bear the greatest proportion of the responsibility of unpaid labour among all four groups. On the other hand, the total mean proportions of time spent in unpaid labour are extremely similar in both developed and developing countries. However, the interaction effect between gender and the development status of the countries is statistically significant. This even though the total amounts of daily unpaid work are the same for both developed and developing countries, women in developed countries do less of time proportionally than their female counterparts in developing countries, we can see how the females in developing countries are experiencing a high level of gender inequality in their societies vis-a-vis their male counterparts.

What the results of this research study have confirmed is the sheer extent of the inequitable distribution of unpaid labour between men and women, as well as its broader implications. Traditionally, in most societies, the breadwinner always has the upper hand over other household members. This thus places female members in a disadvantaged position to their male counterparts in all respects. When women spend almost eight times the amount of time that men spend on unpaid care work, they represent only 35% of the labour force.¹¹ However, when women spend less than two times the amount, their labour force participation increases to 50% of the active population.¹¹ And therefore, women's unpaid care work often remains unrecognized and undervalued, while men receive a larger share of income and recognition for their economic contributions. The evenly distributed unpaid care work between men and women would help in alleviating the gender inequality. Currently, as females are primarily tasked with carrying out various unpaid tasks like unpaid care work, unpaid work for subsistence, or invisible contributions toward market production,⁵ their ability to dictate the quality of their participation and pay is necessarily stymied.¹ Furthermore,
women with higher housework responsibilities are more likely to enter flexible, part-time, or informal jobs, at establishments in the vicinity to supplement the household income. Also, the burden of unpaid care work correlates with higher gender wage gaps. Women earn 65% of male wages when they spend twice as much time as men on unpaid care activities and 40% of male wages when they spend five times more.¹¹ As a consequence, females also face all types of exploitation at work, which undermines their physical, mental, and overall well-being. Finally, because unpaid care work is time intensive, women who juggle both paid and unpaid work are often unable to enjoy their right to rest and leisure.²³

The burden of unpaid labour is even more debilitating for women in developing countries. This phenomenon can be attributed to the fact that Women living in socio-economically backward regions, particularly in the predominantly agrarian societies of developing countries, are compelled to spend long hours doing invisible work due to the lack of investments in basic infrastructure. Therefore, the development of facilities that could bring safe drinking water at the doorstep of every citizen would be a single greatest input which would bring down the large quantum of invisible work for women and girls drastically.²⁵ Similarly, the facility of gas stoves at the doorstep of every household would not only save the hours of females unpaid work of getting firewood for cooking, but also help to reduce various health ailments.²⁴ The lesser the amount of invisible work, the greater the amount of time and energy, which could be diverted to more productive economic development of the household or investments in the future such as education for girls.

Thus, one can see how the socio-economic backwardness of a region can play an instrumental role in trapping women and young girls in their vulnerable position. Within a financially fragile rural household system, racked by poverty, the traditional mindset that forces females to engage in unpaid labour, instead of investing in their education for the future, prevails.²⁵ As a result, the entire family remains stuck in vicious generational cycles of illiteracy and poverty.

More than just an economic issue, the prevalence of this historical bias in the inequitable division of unpaid labour between the genders is also a reflection of the entrenchment of patriarchal culture, which transcends the differences between the development status of the countries. This can certainly account for why the differences between the gender had a far greater statistically significant effect than the interaction effect. Regardless of whether they live in developed or underdeveloped countries, women are still overrepresented among the underpaid and unprotected workers around the world. Despite their contributions to the economy, returns to education are lower for women, gender-based wage differentials persist, and market segmentation and occupational segregation further exacerbate inequalities.²⁶

This inequitable division of unpaid labour is more than just a loss for women and a question of gender inequality. Rather, it is an issue that needs to be embraced by entire nations. Studies have suggested that there could be an almost linear relationship between the invisible work that women perform in taking full care of family burden at home and the men getting paid for his work contribution to society.⁵ Time-use data show how many minutes or hours individuals devote to activities such as paid work and unpaid work including household chores and childcare, and self-care activities.²⁷ Unfortunately, because time-use data is not a regular feature of national statistical systems in many countries, unpaid work is primarily invisible. Even where some time-use data are available, this unremunerated work is not yet recognized as important work that should be covered under economic policies.²⁸ As such, females' unpaid work is reduced to an invisible contribution to the national GDP, which is not acknowledged. Although it is difficult to quantify the value of unpaid work done for others within households, whose quality cannot be reduced to numbers, various studies and models have shown that the monetary value of women's unpaid care work globally for women aged 15 and over is at least USD10.8 trillion annually.²⁸ What this figure suggests is that, without a shadow of a doubt, no nation can develop without the participation of women. There are several ways to close the time gap. Diane Elson has suggested a model with three interconnected dimensions that seeks to address and incorporate unpaid care work into the development agenda: Recognition, Reduction, and Redistribution.²⁹

Therefore, it is long overdue that policymakers redress this imbalance through proactive measures that can transform the lives of women for the better:

● **Development of economic and social infrastructure:**

Economic development goes hand in hand with social development. Technology plays a key role in reducing the time chores. Socio-economic infrastructure in neglected areas could play a prominent role in helping women reduce their quantum of unpaid work. Think about the time when women in many developed regions use to spend on washing clothes and cooking before the invention of modern appliances. Investing in public infrastructure and providing basic services such as childcare can lift the constraints on women's time and increase their participation in the labour force.

On a larger scale, the larger investments in developing social infrastructure like education institutions and health facilities can go a long way in lifting the population below the poverty line and enabling it to benefit from the economic growth of mainstream society.

● **Institution of gender-equitable policies:**

Redistributing more unpaid work to men can be implemented through policies, like paid family leave. Working women would have the opportunity to return to their jobs after having a baby when they have paid leave, while men could be granted paternity leave to spend extra hours on childcare.¹³

● **Free/affordable compulsory education for all:**

Traditional societies in developing countries, like India, particularly in the rural areas, do not support education for females. The prospects of females in improving their life circumstances through education are often limited by the oppressive and conservative attitudes of the male heads of patriarchal households. Illiteracy among females plays a leading role in perpetuating their vulnerability and inequality. Therefore, free, and com
pulmonary state-sponsored education at the grassroots level will provide girls with the opportunity to transcend their situation so that they would have the wherewithal and the confidence to fight for gender inequality when they are grown up. With education, women, certainly those in the developing world, will be able to redress the vast gap in the gender disparity and catch up to their counterparts in the developed world.

- Creation of a model to generate monetary values for unpaid labour:

What could really restore the imbalance in gender disparity, while acknowledging the importance of the unpaid work done by women, is to develop empirical models that could quantify unpaid labour and provide women with a source of income.

The study reveals two key phenomena that should be redressed. First, there is a noticeable difference in the quantum of daily unpaid care work put in by the women in comparison to men across the globe, irrespective of economic development, the difference. Second, there is a far more pronounced male-female disparity in the amount of time spent in unpaid labour in the case of developing economies of the world. Therefore, clearly, the entrenched gender-stereotypical norms of the patriarchal society, coupled with the socio-economic realities of countries, play a dominant role in perpetuating the current gender-related inequality in the amount of time spent in unpaid labour. Therefore, policies pushing for the recognition of unpaid care work by apportioning payments to it and/or the redistribution of unpaid care work between women and men would go a long way in shaping the societies for a better progressive future.

The quantitative aspect of unpaid work imbalance in the light of other factors like education and employment impacting the Economic growth would be an area of future research. Further, the aspects concerning different cultures, ethnicities, and geographical regions can be included in the study.

Acknowledgements

The researcher would like to thank Prof. Maitrayee Chaudhuri for supervising the completion of her project and guiding her through various statistical tools used for the analysis of data. Additionally, I would like to extend my thanks to Kah Ying for helping me organize the research paper better.

References

Communications, 6(111). https://doi.org/10.1057/s41599-020-0488-2


- **Author**

Neasha Mittal is a 17-year-old poet, writer, and author of the anthology ‘Rain Takes the Rainbow by Storm’. A New Delhi, India, based high school student, Neasha actively seeks to empower women with self-sufficiency and financial independence through expanded access to entrepreneurship opportunities and education.

- **Appendix**

Mean Percentages of Day Spent in Unpaid Labour Per Day of Males and Females in Developed and Developing Countries

<table>
<thead>
<tr>
<th>Developed</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>5.7</td>
<td>10.5</td>
</tr>
<tr>
<td>Austria</td>
<td>10</td>
<td>19</td>
</tr>
<tr>
<td>Belgium</td>
<td>10</td>
<td>16</td>
</tr>
<tr>
<td>Canada</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>Denmark</td>
<td>11</td>
<td>16</td>
</tr>
<tr>
<td>Hong Kong</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>Finland</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>France</td>
<td>9</td>
<td>16</td>
</tr>
<tr>
<td>Germany</td>
<td>10</td>
<td>16</td>
</tr>
<tr>
<td>Japan</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>Greece</td>
<td>7</td>
<td>18</td>
</tr>
<tr>
<td>Switzerland</td>
<td>11</td>
<td>17</td>
</tr>
<tr>
<td>USA</td>
<td>10</td>
<td>16</td>
</tr>
<tr>
<td>UK</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>Sweden</td>
<td>13</td>
<td>16</td>
</tr>
<tr>
<td>Poland</td>
<td>10</td>
<td>18</td>
</tr>
<tr>
<td>Norway</td>
<td>12</td>
<td>15</td>
</tr>
<tr>
<td>New Zealand</td>
<td>11</td>
<td>18</td>
</tr>
<tr>
<td>Estonia</td>
<td>11</td>
<td>17</td>
</tr>
<tr>
<td>Ireland</td>
<td>8.8</td>
<td>20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Developing</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mexico</td>
<td>10</td>
<td>30</td>
</tr>
<tr>
<td>India</td>
<td>1</td>
<td>23</td>
</tr>
<tr>
<td>Argentina</td>
<td>9</td>
<td>23</td>
</tr>
<tr>
<td>Brazil</td>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>Armenia</td>
<td>4</td>
<td>22</td>
</tr>
<tr>
<td>Bhutan</td>
<td>6</td>
<td>15</td>
</tr>
<tr>
<td>Bolivia</td>
<td>12</td>
<td>23</td>
</tr>
<tr>
<td>Bulgaria</td>
<td>9</td>
<td>18</td>
</tr>
<tr>
<td>Thailand</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>Cambodia</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>South Africa</td>
<td>6</td>
<td>16</td>
</tr>
<tr>
<td>Turkey</td>
<td>4</td>
<td>19</td>
</tr>
<tr>
<td>Cameroon</td>
<td>5</td>
<td>16</td>
</tr>
<tr>
<td>Chile</td>
<td>10</td>
<td>22</td>
</tr>
<tr>
<td>Costa Rica</td>
<td>8</td>
<td>22</td>
</tr>
<tr>
<td>Kazakhstan</td>
<td>8</td>
<td>17</td>
</tr>
<tr>
<td>Cuba</td>
<td>5</td>
<td>16</td>
</tr>
<tr>
<td>Colombia</td>
<td>5</td>
<td>18</td>
</tr>
<tr>
<td>China</td>
<td>6</td>
<td>16</td>
</tr>
<tr>
<td>Russia</td>
<td>8</td>
<td>18</td>
</tr>
</tbody>
</table>

DOI: 10.36838/v413.12
Applications of Australian Native Aquatic Plants on Purifying Wastewater Sources

Nguyen Nathan
Cavendish Road State High School, Cavendish Rd & Holland Rd, Holland Park, Brisbane, Queensland, 4121, Australia; bachnguyenau@gmail.com

ABSTRACT: Wastewater treatment plants play an important role in maintaining the health of ecosystems and ensuring the economic, social and political soundness of communities. However, current wastewater treatment methods are economically and environmentally unsustainable. Aquatic plant restoration has been receiving attention because of its high efficiency and eco-friendliness compared to previous methods. Three types of aquatic plants: duckweed (Lemna disperma), azolla (Azolla pinnata) and hornworts (Ceratophyllum demersum) were tested for their effects in removing constitutes (pH, total alkalinity, hardness, sulfite, copper, nitrite, nitrate, total N) and reducing the number of bacteria colonies in a wastewater source, over the course of 7 days.

The results show that all three aquatic plants were capable of recovering and removing bacteria in water. Duckweed, however, was the most effective of all three plants. A 3D digital model of a duckweed based wastewater treatment plant was devised to showcase how duckweed could be incorporated into full-scale water treatment systems. The uptake of aquatic-plant based wastewater treatment systems has been slow. The conducted research adds to the advantages and the feasibility of full-scale aquatic-plant based wastewater systems.

KEYWORDS: Earth and Environmental Sciences; Water Science; Wastewater; Aquatic Plants; Phytoremediation.

Introduction

Literature Review:

Wastewater is the used water supply of a community which has been modified physically, chemically, or biologically as a result of the introduction of certain substances such as body wastes, hair, food scraps, fat, chemicals, dirt, and microorganisms. Wastewater is a common source of water pollution which can affect human health and natural ecosystems. Excess nutrients such as nitrogen and phosphorus in a wastewater source, when discharged to rivers and streams, can cause eutrophication. An abundance of nutrients leads to a large ‘bloom’ in growth of algae and phytoplankton, causing an increase in chemical oxygen demand (COD) in the water. As the algae and phytoplankton die, the bacteria consuming them have the potential to remove most of the available oxygen, hampering aquatic biota. The improvement of water quality requires reductions of excess nutrients, total dissolved solids, COD, total coliform numbers, and adjusting the pH, dissolved oxygen, and temperature. This ensures that the water that is disposed or recycled does not present harmful effects.

In Australia, wastewater treatment systems are vital in maintaining the health of the diverse ‘flora and fauna’, and ensures economic, social and political soundness. Moreover, efficient wastewater management addresses the growing issue of water scarcity and allows efficient water recycling in regions with high water stress. Nevertheless, the conventional wastewater management system is economically and environmentally unsustainable. It does not align with basic principles of water conservation, generates toxic sludge and chemicals as by-products, and does not enable effective reclamation and reuse of water and nutrients. Chemical remediation, for instance, may increase the toxicity of the water body, and chemical agents may cause secondary pollution. In recent years, aquatic plant restoration (phytoremediation) has been receiving attention because of its high efficiency and eco-friendliness compared to conventional remediation techniques. Some aquatic plants, such as duckweed and water hyacinth, can store large amounts of inorganic nutrient salts such as nitrogen and phosphorus compounds. Furthermore, Dhir found that aquatic plants can remove a high level (90%) of pathogenic microbes such as Enterococi, Escherichia coli, Klebsiella pneumonia, Pseudomonas aeruginosa, Clostridium perfringens, Staphylococcus aureus, and Salmonella.

The three aquatic plants tested in this research were duckweed (Lemna disperma), ferny azolla (Azolla pinnata) and hornworts (Ceratophyllum demersum). Duckweed is a floating hydrophyte with one solid root that is usually 1-4 cm long. The plant is often elliptic and does not have a stalk except in the bud stage. Duckweed is native to Western Australia, New South Wales, Victoria and Tasmania. Azolla is a native aquatic plant found in the southern states of Australia in still or slow-moving waterways. The plant is reddish-green with several branches arranged on each side and is usually triangular in shape. Hornwort is a submerged and rootless aquatic plant. It has stems that are free branching, reaching a length of up to 60 cm. The leaves are dark green with lengths of 1-4 cm long and two to three times palmately divided.

Hornworts are native to all Australian states except for Tasmania. A decentralised wastewater management system is an approach where each community manages and processes...
wastewater in their locality. Wastewater in a decentralised system would not be sent to a large processing facility or landfill. This allows small-scale and low-cost wastewater treatment facilities to be integrated with urban agriculture. Water flow would remain small, implying less environmental damage. Moreover, the system enables effective and quick water recycling due to being in a neighbourhood level. As a result, this system targets the high water stress issue in rural and small communities. Aquatic plant purification technologies could be incorporated into decentralised wastewater management across a variety of communities.

**Research Objectives:**

While studies have analysed the functions of individual aquatic plants, there has yet been a comparison between those plants in terms of their effectiveness in purifying wastewater samples. Moreover, full scale applications of aquatic-plants-based wastewater treatment plants are still lacking in Australia and in many countries in the world. Therefore, this research aims to: (1) test the ability of three native Australian aquatic plants including: duckweed, azolla, and hornworts for their effectiveness in adjusting pH, total alkalinity, hardness, and removing chemicals in a sample of contaminated water; (2) test the ability of these aquatic plants in reducing the number of bacterial colonies in the sample of water; (3) propose an application of these aquatic plants in wastewater treatment systems by constructing a 3D digital model.

**Research Questions and Hypotheses:**

The research paper addresses the research question: What are the applications of Australian native aquatic plants such as duckweed, azolla, and hornworts on purifying wastewater sources?

This study hypothesized that all three aquatic plants would be able to remove nutrients and reduce the number of bacterial colonies to a high degree. Nevertheless, duckweed would be the most effective in recovering the amount of chemicals and reducing the amount of total coliform numbers in the contaminated water samples. Duckweed could be incorporated as secondary treatment in full-scale wastewater treatment systems.

**Methods**

Water samples were collected from the Bulimba Creek in Brisbane, Queensland. This water is speculated to have undergone only primary and secondary wastewater treatment. Leaves and rocks submerged under a rapid-flowing current were observed, indicating a high nutrient and oxygen level in the water body. The depth of the creek was measured at 30 ± 0.1 cm, and the temperature of the water was 28.0 ± 0.05 °C. Ten 2.5 L plastic containers with dimensions of 21.4 cm (L) x 12.5 cm (H) x 14.5 cm (W) were utilized to grow the aquatic plants. Six cups or approximately 1.42 L of water were collected from the middle of the creek in every container. The containers were rinsed out three times before the final collection to reduce existing contamination.

Each type of plant had 3 set-ups to reduce random error. The experiment also included one control group, which did not contain an aquatic plant. 5.00 ± 0.005 g of *Lemna disperma* and 30.00 ± 0.005 g of *Azolla pinnata* were measured using a 0.01 g increment digital platform scale. When placed in the containers, the *Lemna disperma* and *Azolla pinnata* covered around 90% of the surface water. The *Ceratophyllum demersum* was measured using a ruler and cut to a length of 10 ± 0.1 cm using a scissor. Each container received two strands of hornworts. The ten containers were placed under a shelter in an outdoor and open environment. The temperature ranged from 25 ±0.5 °C to 30±0.5 °C over 7 days. This environment was ideal for aquatic plant growth since the optimum temperatures for temperate submerged aquatics plants are usually between 25 ±0.5 °C and 32 ±0.5 °C (21) Moreover, the experimental set up closely replicated the atmosphere in a constructed wetland while it reduced the exposure to rain and dust, which could affect the reliability of the results.

Levels of total alkalinity, hardness, sulfate, copper, nitrite, nitrate, and total N (nitrate, nitrite and ammonia) in ten water samples were measured using Allora 14-in-1 reagent water test strips. The pH level was measured using a digital pH meter. Measurements were taken every day for seven consecutive days.

The effectiveness of aquatic plants in removing bacteria such as *E. coli* were quantified by growing bacterial cultures in non-selective, nutrient agar plates at 37°C. 1 ± 0.05 mL of a given water sample were collected using a 3.5 ± 0.05 mL sterilized transfer pipette. This solution was then transferred into the petri dish with a lit candle nearby. The lit candle could reduce the number of bacteria getting into the petri dish while performing the experiment. The solution was spread out evenly using a sterilized inoculation spreader. Plastic tapes were used to seal tightly along the side of the petri dish to reduce contamination from the environment. Petri dishes were placed in a dark and warm place to optimize bacteria growth. They were observed every 12 hours, and the final observation was made after 48 hours as bacteria was not likely to grow further after 48 hours. The number of different types of colonies were counted in order to measure the quality of water. Ten different water samples after the experiment and the water sample before the experiment were tested. Since the original and the control water sample had too many colonies of bacteria to count, serial dilution was performed to measure the number of bacterial colonies present. 1 mL ± 0.05 mL of the water sample was diluted with 9 ± 0.05 mL of fresh water. Then 1 ± 0.05 mL of this sample was transferred into the petri dish, making the solution 100 times diluted. New pipettes were used in every step to avoid cross contamination. Three different colonies with distinct colours — white, orange, and red — were characterised as type 1, type 2 and type 3 respectively.

The 3D model of the proposed wastewater treatment plant incorporating aquatic plants was designed using SketchUp. Tools such as 2 Point Arc (A), rotated rectangle, Rectangle (R), Circle (C), Push/Pull (P), Freehand, Eraser (E), Paint Bucket (B) were used to create the model. The house model was designed by Paul Palamra under 3D Architectural Digest.
Results

Removal of Chemicals:

Three aquatic plants showed an ability to recover water constituents to a high degree. The pH level, for instance, indicated a reduction from 7.6 to 7.0 in duckweed, 7.2 in hornworts and 7.3 in azolla (Table 1). The total alkalinity level showed a great decrease in all three aquatic plant species. The initial concentration of 180 mg/L was reduced to 120 mg/L in duckweed, 130 mg/L in azolla and hornworts (Table 1). The level of hardness in the water sample also demonstrated a 60% recovery rate across three aquatic plants. The initial level of 125 mg/L was reduced to 50 mg/L hardness concentration (Table 1). Similarly, the nitrogen level showed a range of around 40 to 60% recovery rate in three aquatic plants, within which the duckweed group had the greatest reduction from 50 mg/L to 20 mg/L. Levels of nitrite and nitrate dropped significantly in correlation with the total nitrogen concentration. Notably, aquatic plants such as duckweed and hornworts showed a complete removal of nitrite and nitrate from the initial water sample. Furthermore, the three aquatic plants removed all 10 mg/L of sulfite and 50% of the 2 mg/L of copper from the initial water sample.

During the growth of the various aquatic plant species over 7 days, the water quality parameters were measured at 8 time points. Data in Figure 1 indicates that duckweed showed the steepest downward trajectory in alkalinity level, pH, total alkalinity, hardness, copper, nitrite, nitrate and total nitrogen. The pH level of the water sample with the presence of duckweed follows a simple linear decrease and reached a plateau after two days of the experiment (Figure 1A). In comparison, data points for azolla and hornworts demonstrate a regressed logarithmic trend in terms of their alkalinity level over time (Figure 1A). This is evident by the high R2 value of 0.9391 for azolla and 0.9417 for hornworts in log-linear regression. Figure 1B shows that the pH level in the duckweed group followed a steep downward trajectory, reducing from 7.6 to 7 and reached a plateau on day 4. The pH level in the azolla and hornworts experimental groups, in contrast, declined gradually over the experimental period, and only reached a plateau on the two final days. Changes in the level of hardness amongst three plants were similar. As shown in Figure 1C, duckweed, azolla and hornwort groups decreased 62% by the end of the experiment, from 130 to 50 mg/L. The duckweed group reached 50 mg/L on day 3, whereas the hornwort group reached the same level on day 4 and the azolla on day 5. The changes in sulfate level is similar in the duckweed and hornwort groups. They removed all 10 mg/L of sulfate after 2 days whereas it took azolla 4 days to recover the same level of sulfate. The duckweed, hornwort and azolla recovered 50% of copper and reached a plateau after 3, 4 and 5 days respectively (Figure 1E). This trend is similar to the changes over time in nitrate level where duckweed recovered 100% of 1 mg/L of nitrate, the fastest amongst three aquatic plants (Figure 1F). On the other hand, azolla was only able to recover 80% of the nitrate in the initial water sample whereas the duckweed and hornwort groups

Table 1: Chemical analysis of the water sample before and after the experiment.

<table>
<thead>
<tr>
<th>Constitutes</th>
<th>Lemna disperma</th>
<th>Azolla jennica</th>
<th>Ceratidia crenatum</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH</td>
<td>Initial (mg/L)</td>
<td>After (mg/L)</td>
<td>Recovery (%)</td>
<td>Initial (mg/L)</td>
</tr>
<tr>
<td>Total alkalinity</td>
<td>180</td>
<td>130</td>
<td>50</td>
<td>180</td>
</tr>
<tr>
<td>Hardness</td>
<td>125</td>
<td>100</td>
<td>25</td>
<td>125</td>
</tr>
<tr>
<td>Sulfite</td>
<td>10</td>
<td>0</td>
<td>100</td>
<td>10</td>
</tr>
<tr>
<td>Copper</td>
<td>2</td>
<td>1</td>
<td>50</td>
<td>2</td>
</tr>
<tr>
<td>Nitrate</td>
<td>1</td>
<td>1</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>Nitrite</td>
<td>25</td>
<td>0</td>
<td>100</td>
<td>25</td>
</tr>
<tr>
<td>Total H</td>
<td>50</td>
<td>20</td>
<td>30</td>
<td>50</td>
</tr>
</tbody>
</table>

Removal of Pathogens:

Three aquatic plants groups indicated a decrease in the number of colonies grown on the culture media. The duckweed group indicated a significant reduction on all three colony types. Figure 2 shows an 87% reduction in colony type 1, 100% in type 2 and 63% in type 3. The water sample containing azolla showed a 77% reduction in colony type 1, 71% in type 2 and 64% in type 3. Finally, the water sample containing hornworts showed a 63% reduction in colony type 1, 77% in type 2 and 58% in type 3. The control experimental group had a similar number of bacterial colonies compared to the initial number of bacteria.
Discussion

Interpretation of Results:

Assimilation of nutrients from water by aquatic plants is an eco-friendly, efficient and cost-effective wastewater treatment approach. The results align with the hypothesis that all three aquatic plants were capable of reducing the amount of nutrients and bacteria in water, but duckweed was the most effective of all. Duckweed was capable of absorbing, on average, the highest number of constituents from the wastewater sample in the shortest amount of time. Moreover, duckweed also removed the greatest number of bacterial colonies amongst all three aquatic plants. Similar observations have been made in a study by Badr El-Din and Abdel-Aziz\textsuperscript{13} where they found that emergent aquatic plants such as hornworts, have strong roots, allowing them to effectively absorb and store nutrients compared to other types of aquatic plants. Constituent levels in the azolla experimental group demonstrated a more gradual decreasing trend, which usually resulted in a plateau at a later time compared to duckweed and hornworts (Figure 1). Azolla, however, removed a larger number of bacteria than hornworts. Notably, standard deviation error bars for the initial and control experimental group overlap for all three colony types (Figure 2). Error bars for type 3 colonies in azolla and duckweed also overlap. These results indicate that the difference between experimental groups is not statistically significant. Nevertheless, similar data for the initial and control groups implies that no significant random errors occurred, boosting the reliability of the research.

3D Model:

A 3D model was devised to propose a duckweed based sewage treatment system to treat household wastewater for surface water discharge. As shown in Figure 3, household wastewater is first passed through an anaerobic baffled reactor (ABR) for advanced primary filtration. It contains a number of upflow sludge blanket reactors, which the wastewater passes over and under. This allows efficient treatment of almost all soluble organic wastewater.\textsuperscript{23} Moreover, it has a simple structure and operation with no special gas or sludge separation equipment.\textsuperscript{23-25} As a result, it could be easily incorporated in decentralized communities or in tropical and subtropical areas of developing countries.\textsuperscript{26} There will be a control point at the effluent of the ABR to monitor the quality of the water. The water is then discharged into a four-chamber duckweed pond. The depth of the pond is suggested to be between 0.6 meters and 1.5 meters since it is suggested as the most suitable for large-scale duckweed wastewater treatment systems.\textsuperscript{27, 28} The hydraulic retention time would be a minimum of 20 days to ensure that the effluent meets acceptable discharge standard.\textsuperscript{27} The floating duckweed mat works in collaboration with aerobic and anaerobic bacteria to purify wastewater sources. Heterotrophic bacteria decomposes organic waste into mineral components (ammonia nitrogen and orthophosphates) that are readily recovered by the duckweed plants.\textsuperscript{29} As a result, the nutrients in the water are reduced for safe discharge to surface water. Duckweed needs to be harvested every 20 to 30 days to prevent the release of nutrients back into the water when the duckweed dies and to allow the fastest rate of nutrient recovery since overcrowding inhibits duckweed reproduction and growth. As duckweed floats on the water surface, harvesting involves netting or scooping duckweed. In large and broader ponds, duckweed can be harvested using a mechanical harvester. Harvested duckweed can become a food source for cultured animals such as fish, ducks, chickens, pigs and ruminants.\textsuperscript{27, 30-34}
Implications:

Nutrients and pathogens removal mechanisms of aquatic plants can help construct a more eco-friendly and efficient wastewater treatment system, which, ultimately, enables effective water recycling for communities experiencing high water stress. The SDG 6 Synthesis Report by the United Nations on Water and Sanitation states that over 2 billion people live in countries experiencing high water stress. In Australia, some areas near the north and east coast experience high (40-80%) to extremely high (>80%) water stress.³⁶ This study proposes that local government agencies can capitalize on aquatic plants (phytoremediation technology) for treating and reusing wastewater in small and decentralized communities. Wastewater treatment technology utilizing aquatic plants is highly cost-effective and efficient compared to previous nutrients remediation technology. On the other hand, the incorporation of aquatic plants in wastewater treatment plants in local communities also poses benefits such as ground water recharge, climate enhancement, and agricultural or landscape irrigation.³⁸ Furthermore, aquatic plants could be incorporated to treat wastewater from agricultural or industrial run-off with low cost and energy consumption.³⁷ Leng et al.³⁸ found that full scale applications of duckweed based wastewater treatment systems exist in USA, Bangladesh, and China. This boosts the feasibility of a duckweed wastewater treatment system in Australia and countries around the world. This study adds to the important contributions of phytoremediation technology to help enable availability and sustainable management of water and sanitation for all individuals in the future.

Limitations

The experimental design has not considered the combination of aquatic plants as an option to purify wastewater sources. A study by Zirschky and Reed³⁹ shows that a mixture of several aquatic plant species can be less susceptible to diseases than a monoculture in a pond. Furthermore, it is beyond the scope of this study to quantify the effect of aquatic plants on purifying a large amount of wastewater influent. This may have led to aquatic plants recovering nutrients in a short amount of time, restricting the identification of clear trends. Figure 1 lacks control data points, reducing reliability. In the bacterial study, it is unknown if all microbes, especially pathogenic microbes, were removed from the phytoremediation method. The bacterial species in type 1, type 2, and type 3 colonies were also not identified. On the other hand, the 3D model did not consider a water recycling system for household sewage. Furthermore, the model did not separate grey and black household water to enable efficient wastewater treatment.

Recommendations

Future studies should take into account:
1. The combination of multiple types of aquatic plants as an option to purify wastewater sources.
2. The relationship between the number of aquatic plants covering the water surface and water purification level.
3. The relationship between temperature and water purification level of aquatic plants.
4. The effect of different water mediums on water purification level of aquatic plants.
5. The separation of grey and black household water in aquatic plant-based wastewater treatment plants.
7. The performance of phytoremediation in comparison with other biological wastewater treatment methods such as microalgae wastewater treatment.

Conclusion

This research aimed to identify the application of three Australian native aquatic plants (duckweed, azolla, and hornworts) on purifying wastewater sources. Based on analysis of the results, it can be concluded that all three aquatic plants were capable of effectively purifying wastewater sources. More specifically, a reduction in values of major water-quality indicators such as pH, total alkalinity, hardness, sulfite, copper, nitrite, nitrate, total N, and numbers of bacterial colonies were seen across three plants. Duckweed was identified as the most effective in recovering constituents and reducing pathogens. A 3D model of a duckweed based household sewage treatment system has been proposed. This could be valuable to future treatment plants designs. Conventional wastewater treatment methods are economically and environmentally unsustainable. Aquatic plants offer an environmentally friendly and cost-effective technology for treatment of wastewater sources. It is proposed that phytoremediation technology could be utilized in decentralized communities with high water stress to allow efficient water treatment and recycling. Overall, this research adds to the advantages and the feasibility of full-scale aquatic plant based wastewater systems.

Acknowledgements

I would like to express my sincere gratitude to Dr. Peter Darben for helping me with my data analysis. I would like to thank Mrs Maria Masselos for her supervision and
encouragement, which helped me tremendously in the completion of this research project. Thanks to Aiko Chung and Vivian Nguyen for proofreading my research paper. My gratitude is extended to Paul Palamara and 3D Architectural Digest for allowing me to use the house model in this research.

References


36. Luo, T., Young, R. S. and Reig P. 2015. Aqueduct Projected Water


**Author**

Nathan Nguyen is a year 11 student at Cavendish Road State High School in Australia. He has a passion for research, particularly in the field of Molecular Biology and Environmental Science. He also founded an international 501(c)(3) non-profit organisation called Sciencious, which aims to bring STEM resources to underprivileged and under-resourced students.
Identifying Factors Affecting the Spread of Variants of COVID-19 by Regression Models to Inform Policymaking

Peter Ma
Beijing No.4 High School International Campus No.2 Xitiejiang HuTong, XiCheng District Beijing, 100022, China, Peterma2020@outlook.com

ABSTRACT: Variants of SARS-CoV-2 continue to put stress on economies and healthcare worldwide. Governments are trying to take necessary steps to slow down the dynamic transmission of variants. In order to find out the potential key factors that drive the spread of variants, like demographics, health, lockdown, and vaccine-related factors, I proposed a study based on machine learning (ML) methodologies to construct predictive models. Those models revealed and ranked possible discriminatory features. What’s more, as stated by The World Bank, since the shortage of vaccines is a major problem for the whole world, lockdown policies remain a significant part of the comprehensive strategy. Thus, I attempted to conduct regression analysis to find out the key pre-lockdown factors that could help maximize the effectiveness of lockdowns. This study seeks to figure out the discriminatory factors of variants and to guide strategies for combating variants based on those findings. Additionally, it provides guidance on the rollout of vaccines and how to maximize the efficacy of lockdowns to slow down the spread of variants of SARS-CoV-2.

KEYWORDS: Systems Software; Machine Learning; SARS-CoV-2; Variants, Spread; Policymaking.

Introduction

The coronavirus disease 2019 (COVID-19), which is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has spread worldwide since the first known case was identified in December 2019.¹ On March 11, 2020, the World Health Organization (WHO) declared it as a pandemic.¹ At the time of writing, almost 200 million global cases have been confirmed and more than 4.23 million deaths have been reported according to the database of Johns Hopkins University.² The pandemic is not only a global health crisis, but has also caused unprecedented economic losses due to loss of life, business lockdowns, tourism disruptions, and reduced trade. It was estimated by the WHO that, due to the economic depression caused by the pandemic, nearly 820 million people are undernourished.³

Currently, there are two major strategies to handle the pandemic. The first strategy is based on the approved vaccines. For example, four COVID-19 vaccines (Moderna, Pfizer, Johnson & Johnson and AstraZeneca) have been authorized by Sweden.⁴ Because of their high effectiveness, they are believed to be significant in halting the pandemic. For instance, the vaccine effectiveness of Pfizer is about 86% among the general adult population in Sweden.⁵ The second strategy is national or regional lockdowns. At the beginning of the pandemic, lockdown efforts and social distancing were the major strategies for governments to handle the spreading of SARS-CoV-2 due to lack of efficient therapeutic treatments and vaccination strategies. Generally, the lockdown strategy works by cutting off the channels of transmission in the pandemic. Admittedly, these efforts mitigated the global health crisis to some extent. However, unprecedented economic losses were caused by those lockdown efforts, like business shutdowns, tourism disruptions, and reduced trade. Usually, most governments, like the U.S., preferred to lockdown nationally or regionally at the beginning of the pandemic to earn time for the development of vaccines. After the majority get vaccinated, governments are expected to handle the pandemic by building up herd immunity and reopening their countries.

Besides those two strategies, other demographic or social factors may curb the pandemic as well. Satyaki Roy applied several supervised machine learning algorithms and found that population density, testing numbers, and airport traffic emerge as the most discriminatory factors which affect COVID-19 infection rates.⁶ Apart from that, Hannah et al. indicated that males are more likely to get COVID-19 than the female group according to meta-analysis,⁷ which implies that gender may be a factor. Also, people with diabetes mellitus tend to have a higher chance to get infected.⁸ All in all, because COVID-19 is relatively new to the whole world, it is still unknown what are the most efficient strategies to halt its spread. Fortunately, these strategies above flattened the pandemic curve successfully. Take the U.S. as an example. Thanks to the lockdown and vaccination, the three hundred thousand new cases per day in early Jan 2021 decreased to six thousand daily in 5 months.⁹

However, most countries in the world are experiencing another resurgence which is driven by variants of SARS-CoV-2. There are four kinds of variants of concern (Table 1). They are believed to be highly transmissible which increases hospitalizations and lead to higher mortality.¹⁰ Thus, the emergence of the new variants poses several new problems to the current preventive strategies for global public health.

The first problem is that we need to create new models to determine how the mutations drive the transmissibility of SARS-CoV-2. In other words, we must examine if the
previously established discriminatory factors, like gender and population density, are still critical in the transmissibility of the variants.

Table 1: Characteristics of four kinds of variants of concerns.

<table>
<thead>
<tr>
<th>Variant</th>
<th>Date of First identification</th>
<th>Location of First identification</th>
<th>Characteristic mutations</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.1.1.7 (Alpha)</td>
<td>Dec 2020⁹</td>
<td>The United Kingdom¹⁰</td>
<td>1. There are 17 mutations in its genome among which 8 mutations form the basis of the 3 vaccines in the UK.¹⁰ 2. 50% increased transmissibility.¹⁰ 3. Increased hospitalizations and mortality.¹⁰</td>
</tr>
<tr>
<td>B.1.351 (Beta)</td>
<td>Dec 2020¹¹</td>
<td>South Africa¹²</td>
<td>1. 50% increased transmission.¹² 2. Greatly reduced susceptibility to monoclonal antibody treatment.¹²</td>
</tr>
<tr>
<td>B.1.617.2 (Delta)</td>
<td>Dec 2020¹³</td>
<td>India¹⁴</td>
<td>1. Increased transmissibility.¹³ 2. Potential reduced susceptibility to monoclonal antibody treatment.¹³</td>
</tr>
<tr>
<td>P.1 (Gamma)</td>
<td>Jan 2021¹⁵</td>
<td>Brazil¹⁶</td>
<td>1. Increased transmissibility.¹⁶ 2. Greatly reduced susceptibility to monoclonal antibody treatment.¹⁶</td>
</tr>
</tbody>
</table>

Secondly, it is significant for us to demonstrate the efficacy of current vaccines in the context of the new model. As shown in Table 1, there are mutations in the variants forming the basis of current vaccines, like the Alpha variant. So, those mutations will likely affect the efficacy of vaccines. For example, the AstraZeneca vaccine was deemed to be only 66% effective against the Alpha variant and 66% effective against the Delta variant after the second dose.¹¹

As a result, the reduced effectiveness will in turn affect our current strategies which are mainly based on the high efficacy of vaccines. One of the questions most countries are facing is whether we need to lockdown again. In May 2021, several European countries started their new lockdown due to the variants at the cost of economic revival while some countries, like Norway, kept opening without regard to the quick spread of variants. Due to the shortage of scientific strategies, all governments are struggling in finding a balance between controlling the pandemic and reopening businesses. Therefore, a new model should be created to evaluate the effects of vaccines and lockdowns on the infection rates of variants to inform lockdown-vaccine-related policymaking.

Thirdly, identifying the pre-lockdown features which could greatly reduce the post-lockdown infected rates is extremely critical for some areas where they have limited number of vaccines and where lockdowns appear to be inevitable. WHO indicates that in parts of the third world, only less than 1% of the populations get vaccinated.¹² The low vaccination rate implies the importance of lockdowns to slow down the spreading of variants. So, national administrations should try to enhance the effectiveness of lockdowns by controlling the meaningful pre-lockdown features. A new model could help with figuring out those beneficial pre-lockdown features.

COVID-19 models are created using machine learning (ML) techniques and are used to predict the spread dynamics of COVID-19 successfully. Zoabi et al. analyzed 51,831 COVID-19 patients to study the effect of demographic factors, like gender, age, and social interactions on the transmissibility of COVID-19 and concluded that close social interactions are critical to the spreading based on ML approaches.¹³ Satyaki Roy et al. applied several supervised ML approaches to examine the datasets of each U.S. state and they determined population density, testing rates, and airport traffic emerge as the discriminatory factors for their forecasting models.⁶ Khan et al. used regression tree, cluster analysis, and principal component analysis (PCA) methodologies in their studies.¹⁴ However, rare ML models for the purpose of predicting the variants of SARS-CoV-2 have been created or reported so far. Therefore, I hypothesize that discriminatory factors influencing the infection rates of the variants of COVID-19 could be identified by ML approaches.

While more and more knowledge about the spreading characteristics of COVID-19 have been investigated, there is no comprehensive study focusing on that of the variants. To solve this gap, I conducted this analysis and attempted to create prediction models utilizing ML technologies. In this study, I made use of datasets collected from the European Union (EU) and the European Economic Area (EEA). There are several reasons why I selected those 30 countries as my samples. The most important reason is that all those 30 countries are experiencing the spread of the four main types of variants. The data was collected from the 40th week of 2020 to the 16th week of 2021. During the 29 weeks, EU/EEA was experiencing a typical period: from the emergence of variants, followed by starting vaccination, then encountering another resurgence caused by variants. However, their anti-epidemic measures were not the same. They were distributing different types of vaccines. Also, some countries avoided re-lockdown while some countries kept lockdowns throughout the entire 29 weeks. Besides, an agency of the EU called European Centre for Disease Prevention and Control is tracking and monitoring the COVID-19 pandemic and building up concrete and accurate datasets.¹⁵ So, my models and analysis would be more convincing if they are based on these official datasets.

In this study, I attempted to achieve the following objectives.

1. Developing ML models to predict the discriminatory factors which drive the transmissibility of variants. I selected several candidates to be examined in this study based on previous studies. All candidates have been demonstrated to be discriminatory factors influencing the previous infection rates of COVID-19. I tried to test to determine if they are still critical in the models of variants.

2. Determining the effects of lockdowns and vaccinations on the infection spread dynamics of variants in the EU/EEA. This will guide further policymaking. If either vaccinations or lockdown measures could be demonstrated to weigh more on controlling the pandemic of variants, the governments could adjust their anti-epidemic measures based on this study.

3. Analyzing the critical pre-lockdown factors which influence the post-lockdown infection rate to guide policymaking. This objective is essential for those countries that do not have enough vaccines and where lockdown is the first option to minimize the new variant cases.

Here I tried to address these issues by using five different ML structures and analyzing their performances. Compared with previous research, this is the first research to evaluate the feasibility of vaccination and lockdown policies for...
controlling the variants of COVID-19 and to guide further policymaking.

**Dataset**

The dataset used in this study contains diverse attributes from 30 countries in the EU/EEA. Most of those attributes were demonstrated to be discriminatory factors which drive the transmissibility of COVID-19 in previous studies. Besides, this dataset is quite unique. Data for different features were integrated from diverse open sources which have been carefully curated. Last but not least, all data has been preprocessed using Pandas, which is a powerful open-source data manipulation tool in Python. This dataset is available on GitHub (https://github.com/mmm-y/covid-19virus.git). I will introduce all features in the following section in detail.

**Features Description:**

- **Population (2020), Male, Female:** Population (2020) is the population in 2020. And Male and Female stand for the male and female population of each country in 2020 respectively. (source:https://www.statista.com/statistics/611318/population-of-europe-by-country-and-gender/)
  - Male%, Female%, Gender: Male% and Female% are the proportion of male and female among the whole population. Male%=Male/Population (2020). Female%=Female/Population (2020). Also, Gender is the ratio between the population of males and females.
  - Urban%: Urban% is the fraction of urban population and total population in 2020. (source: https://data.worldbank.org/indicator/NY.GDP.PCAP.CD)
  - Extreme_poverty: This refers to the fraction of the total population with an income below the international poverty line of 1.90 dollars per day. (source: https://www.kaggle.com/codesagnik/latest-coronavirus-world-tracker?select=owid-covid-data.xlsx)
  - Land_Area: This attribute is the land area in each country (square kilometer) (source: https://www.kaggle.com/codesagnik/latest-coronavirus-world-tracker?select=owid-covid-data.xlsx)
  - Population_Density: Population Density stands for the density of the population and its unit is People/Km²
  - Med_age: This feature represents the median age lifespan of the country. (source: https://www.kaggle.com/tanuprabhu/population-by-country-2020)
  - Health_care_index: This is an estimation of the overall quality of the health care system, including equipment, professionals, doctors, staff and so on. A higher index indicates better quality of the health care system in a particular country. (source: https://www.numbeo.com/health-care/rankings_by_country.jsp?title=2021-mid&region=150)
  - Hospital_beds_per_thousand: This index represents the number of hospital beds per thousand people in the country. (source: https://www.kaggle.com/codesagnik/latest-coronavirus-world-tracker?select=owid-covid-data.xlsx)
  - Diabetes_prevalence: This is measured by the percentage of the population that have diabetes. (source: https://www.kaggle.com/codesagnik/latest-coronavirus-world-tracker?select=owid-covid-data.xlsx)
  - Lockdown_date: The date of imposition of lockdown at the national level between the 40th week of 2020 to the 16th week of 2021. The lockdown date was obtained from the websites of the official institutions of each country.
  - Lockdown_start_from: This attribute shows the difference in days from the 40th week of 2020 to the week of imposition of lockdown at the national level.
  - Total_vaccinations: The data is from Kaggle which represents the number of total immunizations in the country. The owner collected data from Our World in Data GitHub repository for covid-19 every day.²⁰ (source: https://www.kaggle.com/gpreda/covid-world-vaccination-progress?select=country_vaccinations_by_manufacturer.csv)
  - Johnson&Johnson, Pre_Johnson, Post_Johnson: Johnson&Johnson represents the total number of immunizations of the JJN-78436735 by Janssen Pharmaceuticals Companies of Johnson & Johnson. This data is from the same dataset as Total_vaccinations. Pre_Johnson and Post_Johnson represents the total number of immunizations of the Johnson & Johnson’s Janssen COVID-19 vaccine in the country before the lockdown date and after the lockdown date respectively. These two features are collected from Kaggle. (source: https://www.kaggle.com/gpreda/covid-world-vaccination-progress?select=country_vaccinations_by_manufacturer.csv)
  - Moderna, Pre_Moderna, Post_Moderna: Moderna represents the total number of immunizations of the mRNA-1273 by ModernaTx, Inc. This data is from the same dataset as Total_vaccinations. Pre_Moderna and Post_Moderna represent the total number of immunizations of the Moderna vaccine in the country before the lockdown date and after the lockdown date respectively. These two features are collected from Kaggle. (source: https://www.kaggle.com/gpreda/covid-world-vaccination-progress?select=country_vaccinations_by_manufacturer.csv)
  - Oxford/AstraZeneca, Pre_AstraZeneca, Post_AstraZeneca: Oxford/AstraZeneca represents the total number of immunizations of the AZD1222 vaccine by AstraZeneca-Oxford. This data is from the same dataset as Total_vaccinations. Pre_AstraZeneca and Post_AstraZeneca represent the total number of immunizations of the AstraZeneca vaccine in the country before the lockdown date and after the lockdown date respectively. These two features are collected from Kaggle. (source: https://www.kaggle.com/gpreda/covid-world-vaccination-progress?select=country_vaccinations_by_manufacturer.csv)
  - Pfizer/BioNTech, Pre_Pfizer, Post_Pfizer: Pfizer/BioNTech represents the total number of immunizations of the BNT162b2 by Pfizer,inc. and BioNTech. This data is from the same dataset as Total_vaccinations. Pre_Pfizer and Post_Pfizer represent the total number of immunizations of...
the Pfizer vaccine in the country before the lockdown date and after the lockdown date respectively. These two features are collected from on Kaggle. (source: https://www.kaggle.com/gpреда/covid-world-vaccination-progress?select=country_vaccinations_by_manufacturer.csv)

- **Sputnik V, Pre_Sputnik, Post_Sputnik**: Sputnik V represents the total number of immunizations of the Sputnik V vaccines by the Gamaleya Research Institute of Epidemiology and Microbiology in Russia. This data is from the same dataset as **Total_vaccinations**. Pre_Sputnik and Post_Sputnik represent the total number of immunizations of the Sputnik V vaccine in the country before the lockdown date and after the lockdown date respectively. These two features are collected from Kaggle. (source: https://www.kaggle.com/gpреда/covid-world-vaccination-progress?select=country_vaccinations_by_manufacturer.csv)

- **Sinopharm/Beijing, Pre_Sinopharm, Post_Sinopharm**: Sinopharm/Beijing represents the total number of immunizations of the BIBP vaccine, which is inactivated virus COVID-19 vaccines by Sinopharm's Beijing Institute of Biological Products. This data is from the same dataset as **Total_vaccinations**. Pre_Sinopharm and Post_Sinopharm represent the total number of immunizations of the BIBP vaccine before the lockdown date and after the lockdown date respectively. These two features are collected from Kaggle. (source: https://www.kaggle.com/gpреда/covid-world-vaccination-progress?select=country_vaccinations_by_manufacturer.csv)

- **People_fully_vaccinated_per_hundred**: This index represents the number of people who got fully vaccinated per hundred in the country. (source: https://www.kaggle.com/codesagnik/latest-coronavirus-world-tracker?select=owid-covid-data.xlsx)

- **People_vaccinated_per_hundred**: This index represents the number of people who get at least one dose of the COVID-19 vaccine per hundred in the country. (source: https://www.kaggle.com/codesagnik/latest-coronavirus-world-tracker?select=owid-covid-data.xlsx)

- **Total_variants, Pre_variants, Post_variants**: Total_variants shows the total number of variants in the country confirmed in the EU/EEA country. This data is from European Centre for Disease Prevention and Control (ECDC). Pre_variants and Post_variants represent the number of variants confirmed before the lockdown date and after the lockdown date respectively. These two features are collected from ECDC. (source: https://www.ecdc.europa.eu/en/publications-data/data-virus-variants-covid-19-eueea)

- **B.1.1.7_variants, Pre_B_1_1_7, Post_B_1_1_7**: B.1.1.7_variants indicates the total number of Alpha variants confirmed in the EU/EEA country. This data is from the ECDC. Pre_B_1_1_7 and Post_B_1_1_7 represent the number of Alpha variants confirmed before the lockdown date and after the lockdown date respectively. These two features are collected from ECDC. (source: https://www.ecdc.europa.eu/en/publications-data/data-virus-variants-covid-19-eueea)

- **B.1.351_variants, Pre_B_1_351, Post_B_1_351**: B.1.351_variants indicates the total number of Beta variants confirmed in the EU/EEA country. This data is from the ECDC. Pre_B_1_351 and Post_B_1_351 represent the number of Beta variants confirmed before the lockdown date and after the lockdown date respectively. These two features are collected from ECDC. (source: https://www.ecdc.europa.eu/en/publications-data/data-virus-variants-covid-19-eueea)

- **P1_variants, Pre_P_1, Post_P_1**: P1_variants indicates the total number of Gamma variants confirmed in the EU/EEA country. This data is from the ECDC. Pre_P_1 and Post_P_1 represent the number of Gamma variants confirmed before the lockdown date and after the lockdown date respectively. These two features are collected from ECDC. (source: https://www.ecdc.europa.eu/en/publications-data/data-virus-variants-covid-19-eueea)

- **Other_variants, Pre_other_variant, Post_other_variant**: Other_variants indicates the total number of variants except from Alpha, Beta and Gamma variants confirmed in the EU/EEA country. This data is from the ECDC. Pre_other_variant and Post_other_variant represents the number of variants except Alpha, Beta and Gamma variants confirmed before the lockdown date and after the lockdown date respectively. These two features are collected manually from ECDC. (source: https://www.ecdc.europa.eu/en/publications-data/data-virus-variants-covid-19-eueea)

- **Peak_infected_week, Peak_infected_cases**: Peak_infected_cases measures the number of variants at the infected peak. Peak_infected_week shows the difference in weeks from the 40th week of 2020 to the week the peak was reached. Data is collected from the ECDC and rearranged manually. (source: https://www.ecdc.europa.eu/en/publications-data/data-virus-variants-covid-19-eueea)

- **Total_tests, Pre_tests, Post_tests**: Total_tests measures the test number during those 29 weeks. Pre_tests and Post_tests represents the test number before lockdown and after lockdown respectively. (source: https://www.ecdc.europa.eu/en/publications-data/data-virus-variants-covid-19-eueea)

**Data Description:**

A summary statistic description of the above data is shown below (Table 2):

<table>
<thead>
<tr>
<th>Feature</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>0.98</td>
<td>0.035</td>
<td>0.56</td>
<td>1.51</td>
</tr>
<tr>
<td>Urban%</td>
<td>72.73</td>
<td>16.95</td>
<td>15.00</td>
<td>98.00</td>
</tr>
<tr>
<td>GDP_per_capita</td>
<td>43220.77</td>
<td>30410.60</td>
<td>18563.31</td>
<td>181402.83</td>
</tr>
<tr>
<td>Extreme_poverty</td>
<td>1.39</td>
<td>2.82</td>
<td>0.20</td>
<td>15.00</td>
</tr>
<tr>
<td>Population_Density</td>
<td>167.47</td>
<td>253.87</td>
<td>3.00</td>
<td>1386.00</td>
</tr>
<tr>
<td>Med_age</td>
<td>42.73</td>
<td>2.55</td>
<td>37.00</td>
<td>47.00</td>
</tr>
<tr>
<td>Health_care_index</td>
<td>66.88</td>
<td>9.21</td>
<td>52.82</td>
<td>80.56</td>
</tr>
<tr>
<td>Hospital_bed_per_thousand</td>
<td>4.79</td>
<td>1.75</td>
<td>2.22</td>
<td>8.00</td>
</tr>
<tr>
<td>Diabetes_prevalence</td>
<td>6.17</td>
<td>1.81</td>
<td>3.28</td>
<td>9.85</td>
</tr>
<tr>
<td>Lockdown_start_from</td>
<td>75.50</td>
<td>63.11</td>
<td>0</td>
<td>203.0</td>
</tr>
<tr>
<td>Total_vaccine</td>
<td>3.14*10^7</td>
<td>4.10*10^7</td>
<td>5.63*10^4</td>
<td>1.48*10^9</td>
</tr>
<tr>
<td>Johnson&amp;Johnson</td>
<td>5850.36</td>
<td>18934.27</td>
<td>0</td>
<td>99683.00</td>
</tr>
</tbody>
</table>

Table 2: Summary of features and their statistics. The features shown in Table 2 will be used to build the ML models.

DOI: 10.36838/v4/14.14 86
Data Pre-processing:

There are null values for the following attributes: Johnson & Johnson, Moderna, Sputnik V, Oxford/AstraZeneca, Pfizer/BioNTech and Sinopharm/Beijing in Norway and Greece. I replaced the null values with the mean value of the respective columns.

Additionally, when building the ML models, the feature values will be preprocessed by StandardScaler in the Scikit-learn library. The StandardScaler will transform all values so that the distribution of the data will have a mean value 0 and standard deviation of 1. This is required prior to model fitting is because there are diverse variables at different scales, and those variables will cause a bias to the models because of their unequal contributions.

Methods

In this study, I created ML models in the context of the regression algorithm, including support vector regression (SVR), multiple regression (MR), decision tree regression (DTR), random forest regression (RFR), and Bayesian regression (BR). All models were constructed using the Scikit-learn library. The regression problem was to forecast the target label based on existing data. It trained a learner/model based on a training dataset. Then, the model automatically found a corresponding rule to map the input to its output.

Support Vector Regression (SVR): Theoretically, the input dataset to train the SVR model is usually non-linear with respect to its labels (the output). Take our dataset $X$ as an example, it has several input features (N), like the Total vaccine, Urban% and so on. It also has several instances (M), like Norway, Greece and so on. So, $X$ is formed with a $N \times M$ matrix and an output column $y_i$, where $y_i$ represents the total case number of variants. Obviously, it is hard to build a linear relationship between the input matrix ($N \times M$) with respect to $y_i$.

So, how can SVR models help us to dig the relationship rather than the non-linear relationship? Generally, the successfully trained SVR model means to map the input features into a higher dimensional space by a mapping function. There are several mapping functions with the kernel function being the most used. These include polynomial kernel functions, radial basis kernel functions, and multi-layer perceptron kernel functions. Then in the higher dimensional space, we are expected to find a linear regression function $f(x)$ between the projected input and output, and $f(x)$ could predict the output of an unknown object. To be more specific, in our study, $X$ will be projected onto a hyperplane by a kernel function. In this hyperplane, a linear regression function $f(x)$ could be found and used to predict the $y_i$ of an unknown object.

Multiple Regression (MR):

A limit of linear regression is that it can only be trained by a dataset with one dependent variable and one independent variable (One-One structure). However, there are several independent variables, like the Total vaccine and Urban%, in our dataset which forms a N-One structure. So, one of the solutions is to use multiple regression.

Understanding the formula of the MR equation is a good way to illustrate the purpose of MR, which is shown below:

$$X = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_n X_n + \alpha,$$

where $X_n$ denotes the dependent variable, $\beta$'s denote the coefficients for the different independent variables (X's) and $\alpha$ is an error term suggesting the random sampling noise.

As an example, Total variants is our dependent variable as $X_n$ in our study. It probably depends on the first independent variable Urban% as $X_1$, and the second independent variable Diabetes prevalence as $X_2$, and so on. The formula chain will keep going as more and more independent variables are added. Each independent variable has its coefficients to indicate the weight they contribute to the MR equation. So, based on the equation, the objective of the MR equation is to fit a hyperplane to the training dataset based on several independent variables (like the Urban%, Diabetes prevalence and so on) with different weights. As a result, when a new data point is added, we can predict the output (like the number of infection) based on the independent variables of it.

Decision Tree Regression (DTR):

The decision tree algorithm is based on a tree-like structure, which consists of a root node, internal nodes/splits, and terminal nodes/leaves. A typical structure of the decision tree is shown in Figure 1. Generally, decision tree structures could represent a disjunction of conjunctions of constraints on the features of instances. For example, in Figure 1, Li represents ‘constrain 1($f_i < t_1$)’ or ‘constrain 2 ($f_j < t_2$)’ or ‘constrain 3($f_k < t_3$)’, in which this decision tree structure splits a complicated decision into a disjunction of simpler decisions (like constrain 1 or constrain 2 or constrain 3) and each of the simpler decisions (like constraint 1) could be in a conjunction form, like an “and” relationship. When the leaf nodes or our target variables are discrete values, like class attributes, it is called a decision tree classification. And when the leaf nodes are continuous values, like this COVID-19 variants dataset, it is called a decision tree regression.

Basically, how DTR works is to split the space of the original dataset into several sub-spaces by the DTR model. Then, when a new data point is predicted by the DTR, it will assign this new point into the split space to make forecasting. A good example is cited from Georgios Drakos et al. Each of the
instances or data points in the dataset in Figure 2 has two independent variables (X₁ and X₂) and a target variable which are continuous values, although there are only 2 values to simplify the situation. The blue point represents 1 and the red point represents 0. Based on the original dataset, the DTR model was created in Figure 2.a, and the split sub-space is shown in Figure 2.b. The value of each leaf is the average value of each subspace. So, when a new data point, like (0.38,0.5), is predicted by the model, it will be predicted as 0.3333. For my dataset, each instance has five independent features (more details could be found in 5 results) and the target variable is continuous value, which is the number of confirmed cases of variants in each country in the defined time frame. The trained DTR model will split the whole dataset into several sub-spaces (leaves) and the value of each leaf is the mean value of the corresponding subspace. As a result, when a new data point comes, the target variable could be predicted by this trained DTR.

It was shown that the root node contains the entire sample. Each node could make binary or multiple decision to split the dataset within the node into two or more classes. The whole splitting process will keep going until the algorithm reaches the leaf nodes.

**Random Forest Regression (RFR)**

A big problem in the decision tree regression model is overfitting. To address it, I will try random forest regression which uses ensemble learning methods in this study. The simulation of RFR is shown in Figure 3. The RFR will create several independent decision regression models based on subsampling of the entire dataset. The sampling process is conducted at a random and replacement manner. Finally, each decision tree regression will make a prediction and the final prediction is based on the average values of all predictions. In other words, I will select several groups of the training dataset randomly and create more DTR models in the way I discussed in 3.3. So, each model will have an output indicating the predicted number of confirmed cases. The final decision will be made based on the average predictions of all DTR models.

**Bayesian Regression (BR)**

To illustrate BR, some background knowledge is crucial.

**Multiple regression:**

First, let us recall the multiple regression in 3.2. Take our dataset X as an example, X is in the form of \( \{x_1, y_i\} \), where \( x_i \) represents the input variables (like the Urban\% Diabetes prevalence and so on), \( y_i \) is the target labels (like the number of infection), and i is the index of the instance in the X. Then, a linear regression model could be created based on X, and the best model parameters could be set by minimizing the cost function. Finally, it could be used to forecast the predictive value \( w_t \) (like the number of infection), given any input variable \( x_t \).

A mathematical explanation for this linear regression model could be expressed as \( Y = X_0 = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + ... + \beta_n X_n + \alpha \), where \( X_0 \) denotes the dependent variable, \( \beta 's \) denote the coefficients for the different independent variables (X's) and \( \alpha \) is an error term suggesting the random sampling noise. The objective is to find \( \beta 's \) to fit a multiple linear function to minimize the sum of the square error between the predictive value and the target value.

**Bayes Theorem:**

Given the dataset X and each created model represents a corresponding hypothesis, there are various hypotheses that could be created based on X. Our objective is to find the most probable hypothesis (h) fit to X. The Bayes Theorem could be shown as a mathematical formula as well:

\[
P(h|X) = \frac{P(X|h)P(h)}{P(X)}
\]

In this equation, \( P(h) \) is called the prior probability of h, which represents the initial probability that h holds before we have observed X. P(X) is the prior probability that X will be observed. \( P(X|h) \) is the likelihood of the data X given h. We are looking for \( P(X|h) \), which is the probability of h given X and is called the posterior probability.
For my understanding, this Bayes theorem could be organized into:

\[ P(\text{model}|X) = \frac{P(X|\text{model})P(\text{model})}{P(X)} \]

Therefore, I am looking for the model with maximized posterior probability given our training dataset \( X \).

**Bayesian Ridge Regression:**

Then, we are looking for the coefficient vector \( \beta \)'s which minimizes the loss function.

Consider the probability distributions of \( y, y|X, \beta \sim N(\beta^T X - \sigma^2 I) \), where \( y \) is the response feature distribution, \( N \) represents that the conditional is Gaussian distributed, \( \beta^T X \) represents the conditional expectation in ML, and \( \sigma^2 I \) represents the homoscedastic variance. In this equation, \( \beta \) is the fixed unknown vector to be estimated.

Furthermore, if we consider the posterior distribution of \( \beta \) which is also known as maximum a posteriori (MAP) is \( \beta = \text{arg}-\text{max}_{\beta} P(y|X, \beta) \), of which the conditional is Gaussian distributed, \( \beta^T X \) represents the conditional expectation in ML, and \( \sigma^2 I \) represents the homoscedastic variance. In this study, I will create a Bayesian linear regression model based on this equation.

### Evaluation

The evaluation criteria are significant after creating the models. For regression problems, I used two evaluation indicators: the coefficient of determination (R²) and Pearson Correlation Coefficient (PCC).

**Coefficient of determination (R²):**

\( R^2 \) is to determine the percentage variation in the dependent variable that can be explained by other independent variables within a range from 0 to 1.²² It suggests how many data points fall within the regression model. A formula of \( R^2 \) in mathematics is more straightforward.

The formula for \( R^2 \) is²²:

\[ R^2 = \frac{\text{Unexplained Variation}}{\text{Total Variation}} \]

To be more specific, \( R^2 \) represents if the total number of variants (yi) owns a direct relationship to other feature variables (X), like vaccines. If \( R^2 \) is 0%, then it means 0% of the variation in yi can be explained by X or it means none of the data fall within the regression model. Also, if \( R^2 \) is 100%, then it means 100% of the variation in yi can be explained by X or it means all the data fall within the regression model. An ideal model prefers the \( R^2 \) closer to 100%. Additionally, \( R^2 \) can be negative. The negative \( R^2 \) in our model means this model is a poor fit for X or this model cannot set an intercept.²²

Although it is hard to visualize how data is fitted into the regression model on a space more than 3 dimensions, there is a good example by John McDonal to show the relationship between \( R^2 \) and linear regression²³ which can be better understood (Figure 4).

**Pearson Correlation Coefficient (PCC):**

PCC measures how closely the two variables are related. PCC can be either positive or negative to indicate a positive or negative correlation respectively. The absolute value of PCC suggests the degree of correlation with a range from 0 to 1 inclusive (Figure 5). One is the strongest correlation and zero is the weakest correlation (Table 3).

![Figure 4: The relationship between R² and linear regression.](image)

![Figure 5: The relationship between Pearson correlation and linear regression.](image)

It was shown that the data will fit the linear regression model better with a higher \( R^2 \).

Consider there are two sets of data (x column and y column), and each set has n items. Mathematically, we attempt to measure their correlation by the equation below, which means the "covariance of two variables divided by the product of their respective standard deviations"²⁵:

\[ PCC = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}} \]

**Table 3:** The relationship between the degree of correlation with the absolute value of Pearson correlation.²⁶

<table>
<thead>
<tr>
<th>The degree of correlation</th>
<th>[0.8, 1]</th>
<th>[0.6, 0.8]</th>
<th>[0.4, 0.6]</th>
<th>[0.2, 0.4]</th>
<th>[0.0, 0.2]</th>
</tr>
</thead>
<tbody>
<tr>
<td>The coefficient of determination</td>
<td>very strong</td>
<td>Strong</td>
<td>moderate</td>
<td>weak</td>
<td>very weak</td>
</tr>
</tbody>
</table>

The parameter values for all ML models are shown in Table 4. The set of features to create the ML models are Gender, Urban%, GDP per capita, Extreme poverty, Population Density, Med_age, Health care index, Hospital bed per thousand, Diabetes prevalence, Lockdown start from, Total vaccine, Johnson & Johnson, Moderna, Oxford/AstraZeneca, Pfizer/BioNTech, Sputnik V, SinoPharm/Beijing, People fully vaccinated per hundred, People vaccinated per hundred, Peak infected week, Peak infected cases and Total tests. The target value is the number of confirmed cases of variants in each country in the defined time frame.

**Identification of the key factors:**

The objective of this section is to identify the factors that influence the spread of variants. I found all possible combinations (C_{22} = 26334) of any five features as the feature subset from the whole feature set. To avoid bias, every possible feature subset was used to build each ML model 5 times and the PCC and \( R^2 \) will be calculated each time. For example, subset Gender, Urban%, GDP per capita, Extreme poverty and...
Identify the key factors:

Firstly, every feature was transformed into percentiles with respect to their respective column. After that, I sorted all countries in decreasing order based on Post_variants. The top 5 and the bottom 5 countries are shown as a sample in Figure 8.

Then, I selected the top 10 and the bottom 10 countries and used the formula from Satyaki Roy et al. to calculate the weighted average percentile of their features. The formula is shown below:

$$\frac{1}{r} \sum_{i=1}^{r} p(f_i) \cdot (r - p(f_i))$$

where $p(f_i)$ and $p(f_j)$ are the rank and the percentile of the $i$th feature value, and $r$ is the total number of countries in EEA (r=30 here).

I assumed that for the top and bottom 10 countries, the key factors affecting the spread of variants post-lockdown will show the maximum difference in the weighted average percentiles. I plotted the difference in Figure 9. It suggests that Diabetes_prevalence, Lockdown_start_from, Urban% and Pre_tests are the top four key factors.

Population_Density will be used to create an SVR model 5 times with 5 different values of PCC and R². Then the mean values and the standard deviation of those 5 PCCs and R²s will be calculated. For each model, I selected the feature subset which yielded the best performance. The results are shown in Table 5.

Ranking the key factors:

In this section, I applied the extra tree classifiers to rank the features based on impurity. Different features and their impurity are shown in Figure 7. It shows Diabetes_prevalence, Hospital_bed_per_thousand and Lockdown_start_from are the most important three features. Mod_age, Moderna and Total_vaccine follows as their impurities are extremely close to each other. Then, the next echelon are Peak_infected_cases, Pfizer/BioNTech and Total_tests. The last group includes Johnson&Johnson, Sputnik V and Sinopharm/Beijing.

### Table 4: Values of parameters of all ML models.

<table>
<thead>
<tr>
<th>ML models</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVR</td>
<td>kernel = 'rbf', C = 1.0, degree = 3;</td>
</tr>
<tr>
<td>MR</td>
<td>fit_intercept=True; normalize=False;</td>
</tr>
<tr>
<td>DTR</td>
<td>criterion = 'mase'; splitter = 'best'; max_depth=None;</td>
</tr>
<tr>
<td>RFR</td>
<td>n_estimators = 100; criterion = 'mse'; max_depth=None;</td>
</tr>
<tr>
<td>BR</td>
<td>n_iter = 300; alpha_1a=0.6; lambda_1a=0.0000;</td>
</tr>
</tbody>
</table>

### Table 5: For each model, the feature subsets yielded the best result.

<table>
<thead>
<tr>
<th>ML models</th>
<th>Feature Subset</th>
<th>mean(PCC), std(PCC)</th>
<th>mean(R²), std(R²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVR</td>
<td>Med_age, Johnson&amp;Johnson, Moderna, Pfizer/BioNTech, Peak_infected_cases</td>
<td>0.9607, 0.0181</td>
<td>0.9923, 0.0360</td>
</tr>
<tr>
<td>MR</td>
<td>Spunkh V, Total_tests, Hospital_bed_per_thousand, Sinopharm/Beijing, Peak_infected_cases</td>
<td>0.9810, 0.0180</td>
<td>0.9771, 0.0582</td>
</tr>
<tr>
<td>DTR</td>
<td>Med_age, Sinopharm/Beijing, People_vaccinated_per_hundred, Total_tests, Peak_infected_cases</td>
<td>0.9852, 0.0223</td>
<td>0.9957, 0.0572</td>
</tr>
<tr>
<td>RFR</td>
<td>GDP_per_capita, Total_vaccine, Johnson&amp;Johnson, Sputnik V, Peak_infected_cases</td>
<td>0.9945, 0.0177</td>
<td>0.9930, 0.1519</td>
</tr>
<tr>
<td>BR</td>
<td>Diabetes_prevalence, Johnson&amp;Johnson, Lockdown_start_from, People_fully_vaccinated_per_hundred, Peak_infected_cases</td>
<td>0.9905, 0.0324</td>
<td>0.9987, 0.0180</td>
</tr>
</tbody>
</table>

Figure 6: Different combinations of features were used to create different models. Some features were shared by various models.

Figure 7: Ranking the key factors affecting the spreading of variants based on impurity.

Figure 8: A sample of the top 5 and bottom 5 countries based on the Total_variants in decreasing order. For each country, their respective features were transformed into percentiles with respect to the column total.

Then, I selected the top 10 and the bottom 10 countries and used the formula from Satyaki Roy et al. to calculate the weighted average percentile of their features. The formula is shown below:

$$\frac{1}{r} \sum_{i=1}^{r} p(f_i) \cdot (r - p(f_i))$$

where $p(f_i)$ and $p(f_j)$ are the rank and the percentile of the $i$th feature value, and $r$ is the total number of countries in EEA (r=30 here).

I assumed that for the top and bottom 10 countries, the key factors affecting the spread of variants post-lockdown will show the maximum difference in the weighted average percentiles. I plotted the difference in Figure 9. It suggests that Diabetes_prevalence, Lockdown_start_from, Urban% and Pre_tests are the top four key factors.

Figure 9: The difference in the weighted average percentile. The first three features are Diabetes-prevalence, Lockdown-start-from, Urban% and Pre-tests.
Creating the MR model based on the selected features:

In this section, I would like to measure the weight of those four features above based on the MR model and I ignored the least important features from the feature pool. Pairwise Pearson correlations were calculated for each pair of features to figure out if there are any correlated features (Figure 10). From the heatmap, no strong correlation has been found, which indicated I do not need to eliminate any features and I could create the MR model based on those four features. The result of the MR model is shown in Table 6. It indicates that Total_variants has a negative relationship with Lockdown_start_from and a positive relationship with Diabetes_prevalence, Urban% and Pre_tests.

![Image](https://example.com/image.png)

**Figure 10:** The heat-map for each pairwise Pearson correlation. From the

<table>
<thead>
<tr>
<th>Table 6: The result of the MR model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficient</td>
</tr>
<tr>
<td>-------------</td>
</tr>
<tr>
<td>0.0372</td>
</tr>
<tr>
<td>PCC</td>
</tr>
</tbody>
</table>

### Discussion

Nowadays, the variants of SARS-CoV-2 have become one of the biggest problems worldwide. They are threatening our current anti-epidemic strategies. Two major strategies to handle the pandemic are lockdowns and vaccinations. For vaccines of SARS-CoV-2, it was shown that only 15% were fully vaccinated and 29% of the world population had received at least one dose of a COVID-19 vaccine. Unfortunately, the speed of vaccinations is not close to catching up with the spread dynamics of variants of SARS-CoV-2. Also, there are questions concerning the use of old target proteins to produce protective antibodies against those variants. As for the lockdown policy, its objective is to cut off the channels of transmission in the pandemic. However, there are issues with the policy, and the most severe one is the economic depression. Take South Africa as an example, Dorrit Posel et al. estimated that there are around 2.5 million adults in South Africa who lost their jobs due to the lockdown policy from February to April 2020. Another issue besides economic loss is the psychological distress of people in lockdown areas. A systematic review and meta-analysis done by Nader Salari et al. indicates the prevalence of stress in a sample size of 9,074 is 29.6%, the prevalence of depression in a sample size of 44,531 is 33.7% and the prevalence of anxiety in a sample size of 63,439 is 31.9%. So, how to maximize the positive effects of lockdown when lockdown cannot be avoided is our priority. In this study, I tried to address three questions by utilizing computer science.

The discriminatory factors influence the transmissibility of variants:

There are 15 features that have been selected as the key factors that drive the dynamics of spread of variants. They could be organized into 3 categories.

**Vaccine-related features**, like People_fully_vaccinated_per_hundred, Johnson&Johnson, People_vaccinated_per_hundred, Moderna, Sinopharm/Beijing, Pfizer/BioNTech, SputnikV, Total_vaccine.

It showed that vaccines have a close relationship with the spread of variants. Besides, the total amount of vaccines and the proportion of people who have gotten at least one dose of the vaccine could affect the pandemic. This implies the vaccination strategy could still suppress the pandemic, which could be demonstrated by several clinical trials that current vaccines could still offer adequate protection. However, there is no denying that the efficacy of several kinds of vaccines to the variants has been reduced. So, a modified vaccination strategy should be considered. For example, it was reported that some governments, like the U.S., were trying to strengthen vaccine immunogenicity by providing a third dose. Greatly accelerating the speed of vaccinations to ensure more people could be covered under the vaccines plan is another strategy. This is consistent with my research. Even though vaccine-related features are the key factors influencing the spread of variants, they are not the only determinants and other factors are showing more powerful effects which we will discuss in the following sections in detail. Besides, when I attempted to rank those key features, the top three features do not contain vaccine-related features. Thus, to handle the pandemic of variants of SARS-CoV-2, vaccines are one of our priorities. However, the governments should adjust their vaccine strategy to adapt to the new situation with other key factors as well.

**Demographic features**, like Diabetes_prevalence, Hospital_beds_per_thousand, Med_age, GDP_per_capita, Peak_infected_cases and Total_tests.

Among the features, Diabetes_prevalence is the most important. There is evidence that people with diabetes have an increased incidence of COVID-19 and blood glucose control is extremely important for them. In a study by Singh et al., they studied 2,209 COVID patients in China and found 11% of them were suffering from diabetes. And a study processed by Onder et al. found that nearly 36% of 355 COVID-19 patients in Italy were had diabetes. Although the prevalence rates are different in different reports, those indicate patients with diabetes tend to be infected by SARS-CoV-2. Combined with previous studies, I concluded that the prevalence of diabetes is also an important feature for creating the ML model of variants. And I hypothesized that people with diabetes will have a higher chance to infect variants. The biological mechanisms behind this is not clear yet. However, it is reported that the increased glucose levels will elevate the replication level of SARS-CoV-2 through 'the production of mitochondrial reactive oxygen species and activation of hypoxia-inducible factor 1α', which could be demonstrated by a T2DM mice model.
Other potential mechanisms include immunomodulation, renin-angiotensin-aldosterone system, inflammation and insulin resistance. Based on those discoveries, patients with diabetes may be the population most susceptible to variants. So, the government should pay more attention to them when making plans.

Hospital_bed_per_thousand is the second most important feature, which indicates that the spread of variants has a relationship with the hospital capacity. There is other evidence showing the same relationship between SARS-CoV-2 and hospital capacity by using a mathematical model. Dipo Aldila et al. created a “modified susceptible exposed infectious recovered compartmental model” based on the cases from Jakarta, Indonesia. They found that the medical resources, like hospital capacity, are necessary to reduce the burden of COVID. So, hospital capacity may be a crucial feature to slow down the spread of SARS-CoV-2. Admittedly, more investigations should be conducted to find out the possible mechanism. However, I hypothesize that if patients could be admitted into hospitals earlier and if they can stay at a hospital for a longer period, especially during highly contagious periods, the spread of variants of SARS-CoV-2 could be slowed down due to efficient quarantine in the hospital. That suggests one way to reduce the cases of variants is to increase hospital capacity.

Med_age implies that age is as significant a feature for the spread of variants as it is for the SARS-CoV-2. However, a limit in my study is that I only investigated if there is a relationship between the median age and the spread of variants. It is difficult to figure out which age group is more susceptible to the variants. Based on recent CDC reports, it was said that the younger people in the U.S. are taking the place of the older people as the biggest group of newly hospitalized. So, my study could provide a direction for further study of the variants that aims to figure out the age of the susceptible population.

As for the GDP per capita, my study could only demonstrate that the GDP per capita are important features to construct the ML models of variants and indicates that this feature is associated with the spread of the variant cases. This is consistent with a previous study carried out by Shahina Pardhan et al. They demonstrated a negative relationship between the change in COVID-19 cases with the GDP per capita in 38 European countries during the first wave of the pandemic and concluded that the economic performance of a country should be an important consideration for policymakers. However, it is hard to conclude if they have a causal relationship. And if there is a causal relationship, which one is the cause and which one is the result. Because the worsening economic situation may lower the quality of the healthcare system, or the uncontrolled pandemic can hurt the economy significantly.

The Peak_infected_cases indicates that it is possible to predict the total cases of variants by using the number of cases at peak point. In other words, the Peak_infected_cases could be an indicator for the government to evaluate the efficiency of their current strategy in the middle of a pandemic of variants. The lower Peak_infected_cases is associated with a positive output. Thus, the governments could adjust their strategies as soon as possible.

The last feature is the Total_tests. Our study shows the total tests have a relationship with the spread of variants. It is reasonable to deduce their relationship by reviewing the relationship between COVID-19 Cases and testing. Surprisingly, the total number of COVID-19 cases can be reduced with the help of testing. A study conducted by Umit Cirakli et al. demonstrated that 1% increase in the total tests contributed to reducing new cases by 1.45% as analyzed by four models. Before the study, I thought that more new cases of variants will be detected as the outlay of tests, so the total cases of variants will increase. However, based on my study and previous studies, it seems like the spread of variants will be controlled by large-scale testing. The reason is because large-scale testing could detect and isolate risk groups to cut off the channels of transmission. So, governments need to put more effort into large-scale testing.

Lockdown-related features, like Lockdown_start_from. The number of days from our starting point to the day when national lockdowns were carried out is the third important feature in our study. More details will be discussed in section 6.2.

Unimportant features, like Gender, Extreme_poverty, Population_Density, Health_care_index. Those features are deemed to be important for the transmission of COVID-19. Interestingly, in my study, those features are not the crucial factors for the spread of variants. In the later section, I illustrate that the proportion of urban people is important to predicting the post-lockdown infection rate. Combined with the Population_Density here, it seems like it is the density of local areas rather than that of the national level affecting the spread of variants.

Effect of lockdown and vaccination on the infection spread dynamics of variants:

This study shows that both lockdowns and vaccinations are significant strategies to handle the pandemic of variants. Besides, after ranking, I found that lockdown plays a more vital role than vaccinations. That implies the governments should not expect an ideal result by solely depending on vaccinations without any lockdown policies. In addition, my study found that the lockdown length is different in different countries. That’s because countries set the start and end date according to their subjective perspective. Also, no objective criterion existed to guide when they should extend their lockdown date or when they could relieve the lockdown. Furthermore, each country implemented different measures during the lockdowns. For some countries, a national lockdown may simply imply a vacation for schools or working from home. However, for other countries, they may carry out a stricter policy, like quarantining regions.

An important part of the study is to analyze the critical factors which influence the post-lockdown. I constructed a MR model based on four features, Diabetes_prevalence, Lockdown_start_from, Urban% and Pre_tests, and discovered their coefficients. It shows that the coefficient of Lockdown_start_from is -0.251, which demonstrates that the lockdown length has a negative relationship with the spread of variants. Thus, to facilitate a better outcome, governments should lengthen the lockdown length, in other words they should lock down earlier and keep the lockdown period longer. Given that I could deduce the relationship between the lockdown and the
spread of variants based on previous studies of COVID-19 by assuming they are sharing the same pattern, my conclusion could be demonstrated by other studies as well. Based on a study by Atalan et al., they found that the lockdown length during a particular time period without any interruption has a very strong negative correlation with the number of new confirmed cases, with the correlation value being -0.9126. That implies that if the country starts to lock down earlier and keeps it longer, the efficacy of lockdown policy will be maximized.

However, my study only focused on lockdowns nationally. Further studies should concentrate on if lockdowns locally work and what is the criterion for the countries to implement lockdown policies. Because, obviously, lockdowns at a regional scale rather than a national scale would be less harmful. Furthermore, more details about the lockdown could be discussed in a further study, like what kind of measures should be included in the lockdown policy.

Thankfully, vaccinations contribute to the development of the variants’ models as well. Even if they are not the top key factor, eight features out of the fifteen features are vaccine-related features, which suggests that the total effect of all vaccine-related features plays a vital role in the spreading of variants and the vaccines are still useful for the variants. That could guide the creation of policy.

Firstly, a single type of vaccine is far from being able to control the spread of variants and a combination of all types of vaccines is needed. Because each vaccine has different targets, and the mixed use of different vaccines offers better protection for a group of people at risk of being infected by different types of variants. The mixture strategy could be accomplished in two ways. The first one is combining different vaccines in one person. For example, the two doses of Sputnik V vaccine use Ad26 and Ad5 respectively, and its high efficacy has been demonstrated. Besides, John et al. put forward that “using a mRNA vaccine or protein vaccine to boost the first dose of the Johnson & Johnson or AstraZeneca adenovirus vectors could possibly be more effective than giving a second dose of the same vaccine.” Another possible way is to individualize vaccination so that different groups could receive different kinds of vaccines to maximize the performance of vaccines and lower the resistant viruses. For instance, John et al. put forward that it could decrease the resistant viruses by only allowing the younger people to get Johnson & Johnson vaccines. However, this scenario needs to be examined in further studies.

Secondly, in my study, it seems like two features, People_fully_vaccinated_per_hundred and People_vaccinated_per_hundred have equal significance in the spread of variants. That implies some groups of people may not need the second dose or even the third dose to protect them from infection of variants. An investigation by Stamatatos et al. could probably explain that. Their study showed that a single mRNA vaccine dose could boost the antibody level rapidly and significantly in people who have recovered from COVID-19. Of course, more designed trials should be carried out for this question. However, it awakens us to the possibility of giving different groups of people different doses of vaccines, which will significantly relieve the stress from the shortage of vaccines.

**Analyzing the critical pre-lockdown factors which influence the post-lockdown infected rate:**

To identify pre-lockdown factors, I constructed a MR model with a Pearson correlation of 0.8820. The model shows that the prevalence of diabetes, the proportion of population in urban areas, and the total testing before the lockdown are three important key factors that positively influence the post-lockdown infection rate. And the lockdown length has a negative relationship with our target label. Thus, for governments who decide to lockdown their countries, they should increase the testing, reduce the density of urban areas, pay more attention to people with diabetes, and, if possible, lockdown the countries as early as possible and avoid locking down too late.

**The limitations of the study:**

There are some limitations to this study. Firstly, we can enlarge the EEA or include more countries into our dataset, like countries in Africa. The reason why I did not include those in this study is partly because of the incomplete information. The capacity of testing in some countries lags behind the demands of our dataset. In addition, I only identified the key factors affecting the spreading of variants. In further studies, I could analyze each factor in depth. For example, further study could concentrate on weighing the merits and downsides of lockdowns locally or nationally, what is the criterion for the countries to start lockdown policies and what kind of measures could be included in the lockdown policy to maximize its efficacy. Additionally, how to modify the current vaccination strategy to variants is another problem. Currently, some countries, like the U.S., are expected to authorize the third vaccine dose to suppress the spread of variants. However, those policies are not supported by enough scientific evidence.

**Conclusions**

Machine learning is a significant way for governments to predict the spread of variants of COVID-19. By constructing ML models, we could identify and rank the key factors affecting the transmission of variants. Besides, we could figure out critical factors to improve the lockdown effectiveness. While existing studies focus on constructing the ML model of COVID-19, my work presents a possible way to examine that of variants and help guide policymaking. My study illustrates that the spread of variants is affected by multiple factors. Among them, both lockdowns and vaccinations are important to stopping the pandemic. That implies governments should pay attention to every possible factor and not ignore the importance of lockdowns when they are designing the anti-epidemic plan. In addition, special attention should be paid to the people with diabetes and people who live in the urban areas in the blueprint. Individualized vaccination plans are needed, including how many doses of vaccines and what kind of vaccines a person should get. Though this needs to be investigated further. To maximize the performance of lockdowns, the governments, before locking down, should design a strategy based on increasing testing, reducing the density of urban areas, paying more attention to people with diabetes, and, if
some possible solutions or future areas of study. More research should be completed in future studies.

**Acknowledgements**

In writing this research, I have been fortunate to be assisted by my tutor, Dr. Wei Yang, who is the computer teacher in my high school. This research could not have been written without her help. I am deeply indebted to her work to review my drafts and guide my research. As always, I would like to thank my family. For me, I cannot imagine a more supportive family than the one I have.

**References**


**Author**

Peter Ma, Grade 11 student studying at Beijing No. High School International Campus, which is one of the top public international schools in China. Before this, He studied at EagleBrook School in USA, where he got Academic Award issued by the Headmaster and was the leader of the school’s Math team.
Observing Brainwaves when Playing Games Using a Ruby-based Wireless Brainwave Measurement System

Ri He
Takezono High School, Takezono 3-9-1, Tsukuba, Ibaraki 305-0032, Japan; brooknim@gmail.com

ABSTRACT: Using an electroencephalogram (EEG) headset of Mindflex and a Bluetooth module of HC-06, I developed a wireless brainwave sensing system based on the object-oriented program language of Ruby. With the system, the brainwave data from the EEG headset could be wirelessly transmitted to a computer and could be real-time displayed. Using the wireless brainwave measurement system, I measured the brainwaves when playing games and relaxing, and found low-frequency theta wave (0.5 Hz ~ 2.75 Hz) and high-frequency Mid-gamma wave (41 Hz ~ 49.75 Hz) were both increased when playing games.

KEYWORDS: Ruby; Brainwave; EEG; Bluetooth; Mindflex.

Introduction

Many people like playing video games. Game addiction in young people has also become a social problem. It has been observed that video games have both positive and negative effects.¹,² Some educational experts think that violent video games affect the structure and activity of the brain and cause the anti-social behavior in young people.³ However, many scientists and psychologists also find that video games can have many benefits: video games can benefit the mental health of seniors and teach children high-level thinking skills.⁴ To know video games' influence on the brain, I wanted to study the activity of the brain when playing games using electroencephalography (EEG).

EEG is a recording of the electrical activity of the brain from the scalp. It is a non-invasive method to detect the activities of the brain. EEG signals are also called brainwaves. According to the signal frequency, there are five patterns of brainwaves: Delta (δ) waves, Theta (θ) waves, Alpha (α) waves, Beta (β) waves, and Gamma (γ) waves.⁵,⁶ The relations between brain activities and brainwaves are studied.⁷,⁸ A review on wireless sensors for non-medical brain activity measurement was made.⁹ Brain activity while playing computer games was examined using a complex multi-channel EEG system.¹⁰

The frequency of Delta waves is less than 3.5 Hz and occurs in deep sleep. Its amplitude increases when the awareness of the physical world decreases. Theta waves have frequencies of 3.5 Hz to 7.5 Hz and are classed as “slow” activity. Theta waves are strong during internal focus, meditation, prayer, and spiritual awareness. Alpha waves have a frequency between 7.5 Hz and 12Hz. When Alpha predominates, most people feel at ease and calm. Alpha appears to bridge the conscious to the subconscious. Beta waves have a wide frequency range of between 12 Hz and 30 Hz. They are divided into low Beta (12 Hz ~ 17 Hz) and high Beta (17Hz ~ 30 Hz). Beta waves are the dominant rhythm in those who are alert or anxious. It is the state that most of the brain is in when we are listening and thinking during analytical problem solving, judgment, decision making, processing information about the world around us. Gamma waves have a frequency above 30 Hz. They are divided into low-Gamma (30 Hz ~ 40 Hz), mid-Gamma (40 Hz ~ 49.75 Hz), and high-Gamma (> 50 Hz). The gamma wave signal is strong when the brain is at the state of solving problems or high concentration.

In this paper, I will report on a wireless brainwave sensing system using Ruby and the results of brainwave measurements. The brainwave data can be real-time transmitted to and displayed on a computer. Average and normalization methods were used to analyze the data. Ruby is a powerful and easy-learning programming language, very close to spoken languages, and it has libraries of Bluetooth communication, 2D display. Using ruby, the application can be developed in a short time.

Methods

The hardware of the Wireless Brainwave Sensing System:

A Bluetooth module HC-6 and an EEG headset of Mindflex were used to construct the wireless brainwave sensing system. The Mindflex is a brainwave toy letting you control the height of a ball with mental concentration. Figure 1(a) shows the brainwave headset. There are three electrodes in the headset. The sensing electrode is aligned on the forehead. The other two electrodes are clipped to the ear lobes as the reference electrodes. Figure 1(b) shows the circuit board in the headset. Figure 1(c) shows the Bluetooth module of HC-06. The connection method between the HC-6 and the circuit board of the headset was also shown in Figure 1. The default baud rate of HC-06 is 9600. To transfer the EEG raw data, the HC-06 Bluetooth module is set at a baud rate of 57600. Stryger shows the method to change the baud rate of the Bluetooth module HC-06.¹¹

Figure 2 shows the illustration of the wireless brainwave measurement. The voltage between the forehead electrode and ear electrode is amplified and digitalized, then the data is processed by Fast Fourier Transform (FFT) in the headset board to produce the power values of the different brainwave patterns. The digitized raw data and the power values of
the brainwave are transmitted to the computer via wireless connection by Bluetooth.

The EEG headset of Mindflex has two modes: mode 0x01 and mode 0x02. In the default mode 0x01, the circuit transmits at a 9600 baud rate and sends processed data every second: Meditation, Attention, and values for various parts of the power spectrum. To get the raw data of the brainwaves, it is necessary to switch the Mindflex headset to mode 0x02 by sending the hex string 0x00, 0xF8, 0x00, 0x00, 0x00, 0xE0 to the circuit board by the Bluetooth at a 57600 baud rate. In the mode 0x02, the EEG headset transmits the raw data of the brainwaves and the processed values of the 8 patterns of brainwaves.

The EEG headset transmits ThinkGear Data Values, encoded within ThinkGear Packets, as a serial stream of bytes over Bluetooth via a standard Bluetooth Serial Port Profile (SPP). The ThinkGear Data Values includes the following values:

- **POOR_SIGNAL Quality.** It describes how poor the EEG signal ranges from 0 to 200. The higher the number, the more noise is detected.
- **Meters of Attention and Meditation.** The meter values are reported on the relative scales of 1 to 100. A value between 40 to 60 is considered "neutral". A value from 60 to 80 is considered "slightly elevated", a value from 80 to 100 is considered "elevated", meaning it is strongly indicative of a heightened level. Similarly, on the other end of the scale, a value between 20 to 40 indicates a "reduced" level, while a value between 1 to 20 indicates a "strongly lowered" level.
- **Raw Wave Value.** It represents a single raw wave sample. Its value is a signed 16-bit integer ranging from -32768 to 32767. In one second, 512 raw wave values are sent.

**ASIC_EEG_POWER.** This data value represents the magnitude of 8 patterns of brainwaves: delta waves, theta waves, low-alpha waves, high-alpha waves, low-beta waves, high-beta waves, low-gamma waves, and mid-gamma waves. These values range from 0 to 256 and have no units. They are only meaningful compared to each other and to themselves. They are typically output once a second.

The digital data is delivered as an asynchronous serial stream of bytes. The serial stream must be parsed and interpreted as ThinkGear Packets. A ThinkGear Packet consists of 3 parts: Packet Header, Packet Payload, and Payload Checksum.

**Ruby Program for Wireless Brainwave Sensing System:**

Ruby is an easy learning object-oriented program language. It was created by Yukihiro Matsumoto in Japan in the mid 1990s. It was designed for programmer productivity with the idea that programming should be fun for programmers. It emphasizes the necessity for software to be understood by humans first and computers second. I used Ruby to develop the program for the brainwave measurement. Figure 3 shows the flowchart of the ruby program.

The Bluetooth HC-6 was paired with the computer first, and the port of COM3 was set for the communication between the Bluetooth HC-6 and the computer. In the main Ruby program, the ruby library of "rubyserial.gem"
was needed to set up the communication between the Ruby program and COM3.¹³ A ruby library of “ruby2d.gem”¹⁴ was used to display the window and control panel of the brainwave measurement. Figure 4 shows the design of the display window and the control panel developed by using the Ruby program. The raw data and the processed values of brainwaves could be real-time displayed. The Eye-blink light, the Attention light, or the Meditation light would be turned on when the eye blink signal was detected, the attention value was over 70, or the meditation value was over 70.

![Design of the display window and the control panel for brainwave measurements using Ruby.](image)

**Figure 4:** Design of the display window and the control panel for brainwave measurements using Ruby.

The pause-run button and the close button were used to control the measurement of brainwaves. When the save button was pressed, a popup dialog window appeared, and the filename could be chosen. To realize the popup dialog window for filename input, the library of “tk.gem” was needed (Tutorials Point). However, if “tk.gem” was included in the main Ruby program, the speed of the main program became slow, which caused the time-tag of the measurement and brainwave could not be real-time displayed. Figure 5 shows a method to solve this problem. The library “tk.gem” was not included in the main Ruby program <Brainwave.rb>, but included in another program <file.rb>. When the save button was pressed, <file.rb> was then executed and a filename dialog window appeared. The chosen (or inputted) filename was saved in a temp file of <temp.txt>. In the main program and the filename was read from <temp.txt>. Using this method, the speed of the main program was not influenced by the “tk.gem” and the brainwave could be real-time displayed.

```ruby
# Brainwave values
system("ruby file.rb")
file1 = File.open("temp.txt", &:readline)
name = file1.split(" ")
file1.open(name.first, "w") do |f|
  f.puts("savedata")
end
require 'tk'
root = Tk.root.new
root.title = "Window"
names = Tk.getSaveFile
file = File.open("temp.txt", "w")
file.puts(names)
file.close
root.destroy
Tk.mainloop
```

**Figure 5:** Ruby program to pop up the save file dialog window.

---

### Results and Discussion

The image of brainwave measurements is shown as Figure 4. We turned on the power and the EEG headset was worn, and then we ran the Ruby program <Brainwave.rb>, a window appeared shown as Figure 5, and the gathering of brainwave data was achieved via the Bluetooth and was displayed in the window. After measurements, the data was saved by pressing the "save" button. EEG signals were small and susceptible to environmental noise. In order to reduce the influence of this noise, data processing was necessary. Figure 6 shows the data processing method.

First, an unweighted sliding average smooth method was used to reduce the noise of the brainwave data. Second, a normalization method by dividing the value of each EEG pattern by the value of the Low-β wave was used to reduce the measurement error caused by the slight shifts in skin resistance and electrodeposition for each measurement. Then, the average was calculated for each EEG pattern and each person; and finally calculating the average value of each EEG pattern for all the persons.

![Data processing method](image)

**Figure 6:** Data processing method.

For the experimental method, a game of First-Person Shooter (FPS) was used. Since playing games with inexperienced gamers might cause extra tension and stress, the EEGs of 11 experienced gamers between the ages of 16 and 50 years old were measured while they were playing games and while they were in a relaxed state with their eyes closed. Figure 7 shows the normalized average values of each EEG pattern for all 11 persons. From Figure 7, both the high-frequency Mid-γ wave and low-frequency δ wave had larger values when playing the game than those when relaxing.

![Normalized average of each EEG pattern](image)

**Figure 7:** Normalized average of each EEG pattern: Both the high-frequency Mid-γ wave and low-frequency δ wave had larger values when playing the game than those when they were relaxing.
Figure 8 shows the values of Mid-γ waves of the left and right sides of the brain changing with time when playing games. The black line represented the right side of the brain, and the red line represented the left side of the brain. The value of the Mid-γ wave of the right side of the brain was clearly larger than that of the left side of the brain.

**Conclusion**

A wireless brainwave sensing system was constructed using the Ruby program, the Bluetooth module HC-06, and the EEG headset of Mindflex; and brainwave signals were measured when playing games and relaxing.

From Figure 7, both the high-frequency Mid-γ wave and low-frequency δ wave had higher values when playing games than those when participants were relaxing. This result indicated that some parts of the brain were extremely active during playing games that caused the higher value of the Mid-γ wave. In the meantime, the high concentration when playing games might also cause the inactivity of some parts of the brain and the high value of low frequency δ waves.

From Figure 8, the value of the mid-γ waves of the right side of the brain was clearly larger than that of the left side of the brain. This result suggested that the left side of the brain was less active during the game (lower magnitude of Mid-γ wave), while the right side of the brain was very active (higher magnitude of Mid-γ wave).

**References**


**Author**

Ri He is a senior student at the Takemoto High School in Tsukuba, Ibaraki, Japan. He likes science, programming, and music. His research on brainwave measurements started from 2019.
Gap Between Scientists and Lay Public Remains: Genetically Modified Food Knowledge Negatively Influences Opinions

Ryan Hadi
Briarcliff High School 444 Pleasantville Rd, Briarcliff Manor, NY, 10510, USA; ryhadi511@gmail.com

ABSTRACT: Genetically Modified (GM) foods contain ingredients from organisms whose genetic material has been modified with genes from other organisms. Many studies show the benefits GM foods can have for humankind and a majority of scientists agree that there are no known negative health effects caused by the consumption of GM foods. However, there are many misconceptions and misinformation surrounding the topic in the lay public. National consumer polls show that most people think GM foods are harmful to both humans and the environment despite what scientific findings might say, but how much do local consumers know about GM foods? This study assesses the impact of knowledge and attitudes on negative opinions of GM foods. A consent form was sent to residents of a high socioeconomic suburban community in New York State. Those who completed the consent form received the survey, where participants were asked about their opinions regarding GM foods. Descriptive and analytical analysis were conducted on responses using Excel and SPSS, where a multiple linear regression model was developed. Approximately 65% of the sample thought GM foods would lead to health problems for the population, and about 66% of the sample thought GM foods would create problems for the environment. Respondents who claimed to have heard or read most about GM foods thought they were worse for health. The hypothesis was supported by the data which showed that people who live in a higher socioeconomic area say they are more knowledgeable about GM foods are more likely to be subjected to misinformation. The study recommends that community residents should be informed about the misconceptions they have on GM foods and their potential benefits.

KEYWORDS: Social Sciences; Survey; Decision Making; Genetic Engineering; GM Food; GM Opinion.

Introduction
Genetically Modified (GM) foods contain at least one ingredient derived from an organism with an altered genetic composition. Many of the foods Americans eat daily contain ingredients that stem from GM crops, for example, corn starch, corn syrup, and many cooking oils as well as sugar come from GM crops. Of the 12 most common GM crops in the world, nine are grown for human consumption (maize, soybeans, canola, sugar beet, papaya, squash, eggplant, potato, and apples) and three are not (cotton, poplar trees, and alfalfa). GM foods have a promising future because of their health benefits including foods that can be enhanced such as bananas with iron, plants that contain omega-3 fish oils, and grains with increased vitamin A. GM foods were first introduced into the U.S. food market in the 1990s and research to date shows that they are just as safe as non-GMO foods. GM crops have led to great economic benefits at the farm level, where a net gain of 186.1 billion dollars has occurred since 1996 to 2016. Both developed and developing countries have seen equal gains from the use of GM foods. Scientists are in consensus about the potential of GM foods to provide significant benefits to humankind.

There have been and will continue to be significant economic gain produced by genetically modified crops. Sixty-five percent of these gains have come from the sheer number of crops produced, while the other 35% is from cost savings from the use of pesticides and other excess farming aids that are reduced by GM foods. In South America, farmers have been able to spend the time they would usually use preparing crops to harvest an additional soybean crop after wheat in the same season. Even since GM crops have been in use, farmers have been able to reduce the damage caused by pests and weeds.

A study by Qaim and Kouser conducted a comparison farm household survey of Indian farmers who grow Bt cotton and traditional cotton between 2002 and 2008. Bt cotton contains genes from a soil bacterium called Bacillus thuringiensis which gives the cotton a built-in pesticide against the cotton dense bollworm that destroys cotton crops. Scientists found that Bt cotton growing households consumed more calories than traditional cotton growers which suggests that GM Bt cotton adoption helps farmers become more food secure. Bt cotton producers are also able to afford more nutritious foods.

The researchers concluded that this technology reduced food insecurity by 15-20% among GM Bt cotton producing homes. Not only do GMOs allow farmers to gain more profit, but they can increase the global food supply as well. Another useful genetically modified food is Golden Rice, fortified with beta-carotene, a precursor for Vitamin A that could prevent blindness in children. After waiting for a long time, Golden Rice received approval in the Philippines despite a political and regulatory blockade. Golden Rice is just one of many GM crops waiting for its chance to be put into use and benefit humanity.
Many Americans still oppose GM foods based on inaccurate information.⁸ Due to the misconceptions of GM foods' potentially harmful effects, they have been the target of controversy. Those in support of the use of GM foods claim they provide enhanced quality and quantity of foods, they can be produced easily and, in all seasons, to feed the growing human population, and GM foods can reduce the use of pesticide and herbicide use, allowing farmers to save money and make food safe from chemicals. However, those opposing the use of GM foods say there is not enough research on the potential negative health and other effects including the genetic modifications of organisms may speed up natural selection and disrupt the ecosystem, leading to a rise in herbicide resistant weeds and pesticide resistant insects.

Misinformed public opinions lead to policies that can hinder the development of biotechnology of GM crops that remain a useful asset for worldwide development. Debate over the uses of GMOs has intensified and states are now demanding GMO products be labeled as such. This is partially motivated the federal labeling requirement, as to avoid variations in regulations across states. Negative public opinions could have a downturn effect on investments and the future of GMOs, and it is confusing considering scientists consistently posit that GMOs are safe and can provide many benefits to humankind. There is a major gap between the public understanding of the real benefits of GM foods. Only 37% of U.S. consumers agree with the 88% of members of the American Association for the Advancement of Science that GMOs are safe to eat.⁹ A previous study by Scott et al.¹⁰ about attitudes towards GMO foods in Europe and the U.S. found that life scientists have a positive opinion as compared to lay people. Lay people think that the negatives outweigh the positives associated with GMOs. The key finding of this study was that the lay public opposed GMO foods on a moral basis and considered nature and naturalness as sacred. USDA Organic and the Non-GMO project are dominating the market for organic foods. Bain and Selfa¹¹, in their study of non-GMO and USDA organic label found that both these groups are competing for legitimacy. Non-GMO labels want to test for signs of GMOs in organic foods to prevent contamination, however the USDA organic label wants to get rid of GMOs entirely because it thinks that contamination is inevitable. Both these companies exclude GMOs because there are many laypeople who believe GMOs are harmful in general.

Legge and Durant¹² assessed public opinions and attitudes towards GM foods in the European Union. The researchers found that a risk assessment knowledge is not enough, and people are influenced by social, political, and cultural factors when deciding whether to choose GM foods for consumption. The study found that people with a stronger background in sciences and those who trust their government supported the use of GM foods more than those who do not. Citizens are not just emotional but can judge for themselves the costs and benefits of GM foods if informed.

Despite GM technology being decades old and its impact studied extensively by scientists, there are many common public misconceptions around foods containing GMOs, such as they are harmful to both humans and the environment. The goal of this study is to understand the prevalence of negative attitudes surrounding GMOs in a suburban community. National consumer polls conducted by the Pew Research Center¹³ show that most people think that GM foods are harmful to both humans (40%) and the environment (49%) despite what other scientific findings might say. The local attitudes towards GM foods is limited and national surveys do not assess whether these attitudes are influenced by how much respondents have heard or read about GMOs. This study aims to assess the relationship between knowledge and opinion in a local community of New York State.

Hypothesis:

People who have heard or read and care more about genetically modified foods will have a negative opinion regarding GM food potential.

Methods

Sample Description:

According to the U.S. census bureau,¹⁴ as of 2018, the population of the town of Briarcliff Manor is 8,215. Briarcliff school district guardians will represent a large portion of the total population of the town. About 48.5% of the town's residents belong to the age group of 25–64 which translates to 3,984 people. This is the age group that would most likely include the parent population of Briarcliff Manor. Data was collected via convenient sampling methodology, which limits generalization. An email blast to the Briarcliff community was sent via the School District email system.

Survey Procedure:

First to obtain consent, all participants were required to fill out the informed consent form sent in the introductory blast email from Briarcliff Administration that explained the purpose of the study. Once the informed consent form was received, participants, both male and female parents were emailed the survey individually via Microsoft Forms. The survey posed minimal risk to the participants of the study and asked the participants to reveal their dietary practices, education, and income. Participants could opt out of answering questions without penalty at any time should they have experienced any distress. Questions from a validated survey from the Pew Research Center were utilized and adjusted to the specific needs of this study. The responses collected were calculated as aggregate data and the surveys were unlinked from the emails. The Microsoft Form survey data was collected and downloaded to an Excel file. The Excel file was downloaded to SPSS where both descriptive and analytical tests were conducted using multiple linear regression models.

Statistical Methods:

To understand the relationship between knowledge (heard or read) and care (independent variables) and negative opinion (dependent variable) a regression model was developed. The research question was, does knowledge and care predict negative opinion of GM foods? In this regression model equation, the dependent variable is negative opinion scale, which was developed by summing the variables GM bad for health and environment.
Results

Table 1: Sample demographics.

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Number (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender Male</td>
<td>14 (13%)</td>
</tr>
<tr>
<td>Female</td>
<td>66 (67%)</td>
</tr>
<tr>
<td>Average Age</td>
<td>51.92 Std Dev = 9.28</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
</tr>
<tr>
<td>Hispanic or Latino</td>
<td>4 (5%)</td>
</tr>
<tr>
<td>Not Hispanic or Latino</td>
<td>70 (95%)</td>
</tr>
<tr>
<td>Race Black/African American</td>
<td>0</td>
</tr>
<tr>
<td>White</td>
<td>66 (67%)</td>
</tr>
<tr>
<td>Native American or American Indian</td>
<td>0</td>
</tr>
<tr>
<td>Asian/Pacific Islander</td>
<td>6 (8%)</td>
</tr>
<tr>
<td>Other</td>
<td>4 (5%)</td>
</tr>
<tr>
<td>Income $50,000-$100,000</td>
<td>4 (5%)</td>
</tr>
<tr>
<td>$100,000-$200,000</td>
<td>19 (26%)</td>
</tr>
<tr>
<td>$200,000-$500,000</td>
<td>28 (38%)</td>
</tr>
<tr>
<td>Prefer not to say</td>
<td>22 (30%)</td>
</tr>
</tbody>
</table>

There is a discrepancy in the perception of genetically modified food safety between scientists and the lay public. This study aimed to find the relationship between knowledge and opinion regarding GMOs. It was hypothesized that people who live in a higher socioeconomic area will claim to be more knowledgeable about GM food potential are more likely to be misinformed. As per Table 1, a majority of the survey respondents were female (87%) with an average age of 52 years old, 87% of the sample was white followed by 8% who were Asian, and 5% other. About two thirds (64%) of the sample had an income greater than $100,000 and 30% preferred not to say.

GMO Knowledge:

Figure 1: Sixty four percent of survey respondents heard or read a little about genetically modified foods, whereas 34% of respondents had heard or read a lot about genetically modified foods, and only 1% heard or read nothing.

In terms of respondents’ GM food knowledge, this sample population was very knowledgeable (98%) with a majority hearing or reading a little or a lot about GM foods (Figure 1).

Most of the sample respondents (80%) cared about the issue of genetically modified foods (Figure 2). This is not a surprising finding considering most of the respondents have a higher education level.

Respondents believe GM foods will lead to more affordable food (56%) and increase the global food supply (71%) (Figure 3). Some positive messages about GMOs did reach this sample population because they agree with the benefits of GM foods to increase the global food supply and lead to more affordable food. On the other hand, even though scientists across the world know there is no negative health effects caused by GMOs, respondents believe that GMOs are bad for their health.

Figure 3: More than half (56%) of the sample said that genetically modified foods were fairly or very likely to lead to more affordable food. A majority of the study (71%) thought GM foods would increase the global food supply. 66% of the sample thought GM foods would create problems for the environment. Exactly 65% of the sample thought GM foods would lead to health problems for the population.

Figure 4: 90% buy GMO Free foods, and 90% read labels to check for GM ingredients.
Questions on shopping experiences revealed that a majority of community residents buy GM free foods (90%) and read labels to check for GM ingredients (90%) (Figure 4). Respondents avoid GM foods proving their lack of trust in the food industry and GM food.

Figure 5: Most of the respondents did not trust elected officials (73%) nor did they trust the food industry (86%). A majority of the respondents trusted scientists (84%) and small farmers (76%). Trust in the media was split.

According to Figure 5, 84% of respondents trusted scientists. Respondents lack trust in elected officials (73%) and the food industry (86%) unlike the findings of European populations. Even though respondents are wary of the food industry, if more efforts are made by scientists to inform the public about the benefits of GM foods might help alleviate concerns.

Table 2&3: The estimated regression model equation output for the impact of GM knowledge and concern on negative GMO opinion.

The regression model explained the relationship between knowledge and negative opinion. In the equation, the dependent variable (DV) is the negative opinion scale, which was developed by summing the variables ‘GM bad for health and environment’, and the model tested if the independent variables (IV), ‘GM heard or read’, and ‘GM care’ predicted the DV (Table 2). The R or the multiple correlation coefficient is .607 that indicates a large level of prediction. Also, R square which is the proportion of variance in the DV that can be explained by the IV is .368 or almost 37% of the variance in the negative opinion can be explained by the IV, ‘GM Heard Read or Care’. Table 3, the ANOVA table tests shows that the overall regression model is a good fit for the data. The table shows that the independent variables statistically significantly predict the dependent variable at p-value less than .000. In Table 3, unstandardized coefficients show that the negative opinion varies with ‘GM Heard Read’ and ‘Care.’ The unstandardized coefficient, B, for ‘GM Heard ReadLot’ is equal to 1.039, and ‘GM CareALot’ is 1.905, and causes a one point increase in the negative opinion scale.

The regression model was statistically significant and supported the hypothesis that individuals who heard or read of GMOs and personally cared about them held a negative opinion on the subject.

Discussion

Most of the food consumed by humans today, contains some genetically modified ingredients. There is consensus among scientists that genetically modified foods cause no harm to human health as far as our understanding. GM crops can be significantly beneficial for increasing the global food supply, as well as, making food more affordable. GM foods have a great potential to eliminate hunger and poverty, particularly in developing nations. Misconceptions around GM foods are prevalent in American society, where people believe contrary to scientific knowledge that GM foods are in fact detrimental to their health and environment. This misunderstanding has grave consequences, as GM food research and production are constantly being held back for political and ethical purposes by GM food opponents. In addition, non-GMO food organizations and organic food labels have demonized GM foods.

This study’s hypothesis was supported with evidence from the data that people who claim to have heard or read more about GM foods and personally care about them hold a more negative opinion on the issue. This is explained by the anti-GM movements who make GM foods appear worse than they really are. It is important that this misunderstanding is cleared, and scientists continue to inform the public on the benefits of GM foods.

Some limitations of the study are that the sample was mostly female participants who resided in a higher socioeconomic area. I don’t know if the results would have been the same had there been more male participants. Also, the population of Briarcliff Manor is about 7,616 and we only received 75 surveys, so our margin of error is 11% indicating a slightly low response rate. It is important to note that people who strongly oppose GM foods may be keener to represent...
themselves as well-informed on the topic than those who are neutral so the correlation between negative opinions and how much each subject heard or read and personally cared might be slightly overstated.

Future study will aim to survey and compare participants of a more diverse socioeconomic area and that includes more men. The next phase of my study aims to compare the GM knowledge among residents of a lower-socioeconomic area, in addition to identifying the sources and types of misinformation that perpetuates negative attitude towards GM foods, in both areas.

Acknowledgements
This study would not have been completed without the support of my science research class teacher, Mrs. Melissa Carnahan, the school administration at Briarcliff High School including Principal Deborah French, Vice Principal Diana Blank, and other science research teachers. The science research teachers provided the direction for my day-to-day research project. I appreciate the support of the IRB who made me think about the privacy of the survey participants. My mentor, Dr. Philip Fernbach from the University of Colorado, who is a renowned expert in cognitive science and decision-making took the time out of his busy schedule to provide feedback on my study. Lastly, I would like to thank my family for their everlasting support of my research endeavors.

References

Author
Ryan Hadi is a rising senior at Briarcliff High School, in Briarcliff Manor, New York. He is passionate about his research on the issues surrounding GMOs, which he conducted over two years. He presented his research findings various science competitions including the Westchester Science and Engineering Fair, where he won first place and was nominated to participate at the Genus Olympiad international Science Fair, where he went on to win a bronze medal. To address the misinformation around GMOs found by this research, Ryan followed-up with his survey respondents and shared Cornell Alliance flyers clarifying misconceptions.
Controlling the COVID-19 Pandemic with Masks and Social Distancing

Shuying Liu
Kunming No.1 High School, 1 Kunhua Road, Kunming, Yunnan Province, People's Republic of China; Shuyingliu05@hotmail.com

ABSTRACT: Coronavirus disease 2019 (COVID-19) is a severe acute respiratory infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The pandemic has spread to 223 countries and regions around the world, and the total number of infections has exceeded 211 million, with over 4 million deaths. COVID-19 is causing substantial social and economic losses, while drugs and vaccines are still in the developmental stages or needing full evaluation. Therefore, some non-pharmaceutical interventions (NPI), such as wearing masks, social distancing, and isolating at home, are currently the most important measures for controlling disease spread. This article reviews the role and effects of wearing masks and maintaining social distancing in the control of COVID-19.

KEYWORDS: Translational Medical Sciences; Disease Prevention; COVID-19; SARS-CoV-2; Mask; Social Distance.

Introduction
Coronavirus disease 2019 (COVID-19) is a severe acute respiratory infection caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). As of August 22, 2021, the disease had become a pandemic, spreading to 223 countries and regions around the world, with a total of more than 211 million infections and approximately 4.4 million deaths.¹ In response to this global crisis, the World Health Organization (WHO) issued response guidelines.² Meanwhile, to control the epidemic to the greatest extent, each country introduced relevant policies based on their own conditions, endeavoring to stop the increasing of mortality rate and reduce the socioeconomic burden caused by COVID-19.

SARS-CoV-2 is a member of the coronavirus family (Coronaviridae). The virus particles are mostly round with a diameter of 80-120 nm.³ Particle envelopes are covered with corona-shaped spike proteins on the periphery, hence the name. The virus can be alive and infectious under cold conditions but easily inactivated by heat. Moreover, it can survive for 4–72 hours on a variety of objects, such as plastic, stainless steel, cardboard, and other surfaces. The main routes of transmission are droplet and contact transmission.⁴-⁶ COVID-19 caused by SARS-CoV-2 mainly manifests as fever, dry cough, and fatigue. A small number of patients have nasal congestion, runny nose, diarrhea, and other upper respiratory and digestive tract symptoms.⁷,⁸ Severe cases often cause difficulty in breathing after 1 week and rapidly progress to acute respiratory distress syndrome, even death. Most people with infections have a good outcome, especially children with mild symptoms, but elderly individuals and those with chronic underlying diseases can have more severe symptoms.

Discussion
The Transmission of SARS-CoV-2 is Enhanced by Personal Behavior:

At present, it is known that the transmission of SARS-CoV-2 mainly occurs through droplet and contact transmission. In some specific aerosol production processes or environments, it may be transmitted via aerosols.⁹ The incubation period of COVID-19 is 2–15 days, with an average of 5.1 days. Patients with asymptomatic infections and those in the incubation period are highly contagious.¹⁰,¹¹ Studies on the dynamics of SARS-CoV-2 spread show that patients in the incubation period are capable of transmitting virus on average 2.3 days before symptom onset, and the intensity of infection reaches a peak 0.7 days before obvious symptoms.¹² A large viral load appears in the saliva of patients and asymptomatic infected persons and spreads through droplets formed by coughing, sneezing, talking, breathing, and other behaviors. Although infected children are mostly mildly ill, there is no significant difference in the viral load in their saliva compared to that of adults or between those who are asymptomatic and symptomatic. Statistical studies have shown that among all SARS-CoV-2 infections, 40–45 % of the sources of infection are from asymptomatic infections.¹³

Epidemiologically, the transmission ability of a pathogen or virus is expressed by the basic reproduction number (R0). Referring to the situation in which no external force is involved, and everyone is assumed to have no immunity, the R0 number is the average number of people infected by a person with a certain pathogen. The larger the R0 number, the more difficult it is to control the epidemic. If R0<1, the infectious disease will gradually disappear; if R0=1, the infectious disease will become endemic in the population; and if R0>1, the infectious disease will spread exponentially and eventually become a pandemic. It is estimated that the R0 of SARS-CoV-2 is 2.4–3.9.¹⁴ For such a highly infectious disease, the goal of prevention and control is to employ medical and nonmedical methods to reduce the R0 to less than 1.
Nonpharmaceutical Interventions (NPIs) to Control the COVID-19 Pandemic:

Although COVID-19 vaccines are currently being deployed on an emergency basis throughout China, the United States, the United Kingdom, and other countries, people are now facing different variant stains of SARS-CoV-2, and vaccine efficacy needs more evaluation. Therefore, the use of nonpharmaceutical interventions (NPIs) to reduce COVID-19 infections is particularly important at this stage, even over a long period of time. Although home isolation, community restrictions, and school closures are most effective measures during a pandemic, they also bring about a series of negative social and economic effects. Thus, some other nonmedical interventions that take into consideration various implementation convenience and economic factors, such as wearing masks and maintaining effective social distancing, are currently the most recommended WHO Preventive Measures and can effectively reduce the risk of virus transmission from potentially infected persons.²

Masks:

The use of masks can be traced back to the 14th century when people realized that covering the nose and mouth could block the spread of respiratory diseases. By 1910, the Chinese doctor Wu Liande designed the "Wu-style mask" during the plague epidemic, which played a vital role in protecting the population. His concept was known as "a milestone in modern plague epidemic, which played a vital role in protecting the population. His concept was known as "a milestone in modern preventive measures".³ Since then, East Asia, including China, Japan, South Korea, and other regions, have gradually developed a "mask culture". During the high incidence of respiratory viruses in autumn and winter each year, many people wear masks spontaneously to prevent disease infection. However, due to cultural differences, this behavior in the eyes of Europeans and Americans is tantamount to declaring that I am a "patient" and that I am the source of infection on the street. Therefore, many healthy people refuse to wear masks to protect themselves and prevent disease infection. This difference in perception in the initial stages of the COVID-19 pandemic in 2020 meant that countries in Europe and the United States saw a surge of infections. Studies have shown that in countries and regions that do not require mandatory wearing of masks, the spread of SARS-CoV-2 is 7.5 times that of requiring mandatory wearing of masks.⁴ Comparing the incidence of COVID-19 per million people in Hong Kong, mainland China, Spain, Italy, and Germany in March 2020, we found that in Hong Kong, China, where 96 % of the population wore masks, the number of COVID-19 cases was only 129 per million people. In the same period in European countries without a mask culture, the number of COVID-19 cases per million population was 1,242 cases in Germany, 2,251 cases in Spain, and 2,983 cases in Italy, which was approximately 9.6–23.1 times higher than that in China.¹⁷

From medical databases, Chu and his colleagues used meta-analysis to review medical masks used in medical units in the COVID-19, SARS (Severe Acute Respiratory Syndrome) and MERS (Middle East Respiratory Syndrome) seasons. They found that in the group that does not wear medical masks, the probability of contracting the virus is 17.4 %, while in the group wearing medical masks, the probability is only 3.1 %, which shows that medical masks can do a good job of personal protection in a high viral load environment. This protective effect is even greater when wearing N95 masks. Wearing a mask can not only reduce the direct transmission between people due to droplets but the physical barrier can also effectively limit the virus that is transmitted by potentially infected persons when they speak, cough, or sneeze, spreading emission clouds of viral particles and contaminating the surface of objects. In addition, because of wearing a mask, the diffusion range of emission-containing virus particles can be reduced from 2.4 m to 6 cm.¹⁸

However, some negative effects of wearing a mask for a long time cannot be ignored, such as the influence of facial recognition and emotional communication between people, the influence on talking, especially communications with the hearing impaired, the discomfort people have when wearing glasses that fog up, and for people with diseases that cause breathing difficulties. Moreover, crowd psychology in populations choosing to wear masks or not need to be considered when implementing mandatory mask-wearing measures. At present, more than 100 countries around the world have adopted the mandatory wearing of masks as a compulsory public health measure at the national level. Investigations have shown that if only voluntary mask wearing is recommended, it will lead to a weak or even ineffective epidemic prevention outcomes and even cause negative effects, such as social injustice.

Social Distancing Supports the Protective Effects of Masks:

Maintaining a certain social distance is of great significance in preventing infectious respiratory diseases, especially those that are spread by droplets. Droplet particles usually fall within 1–2 m, depending on the size of the droplets. Keeping effective social distancing can reduce the probability of transmission of respiratory viruses. A meta-analysis on confirmed cases involving COVID-19, SARS, and MERS revealed that when the social distance is less than 1 m, the probability of viral infection or transmission is 12.8 %; however, when the distance is more than 1 m, transmission probability is reduced to 2.6 %. Meaningfully, every additional meter of social distance can reduce the probability of infection by 2.02 times.²⁹ Therefore, the social isolation strategy currently adopted by China is one-meter space between persons, while in Western countries, it is 2 meters (approximately 6 feet) between people. However, a long time spent maintaining social distance, especially social isolation, may also bring certain negative effects, including psychological problems such as depression and anxiety.

Nevertheless, maintaining social distance can enhance the effectiveness of wearing masks to control epidemics. A mathematical statistical model shows that effective social distancing supported by related policies can reduce the number of hospitalized cases of COVID-19 patients by 92 % in the United States, which is very important for saving medical resources. Their predictions also show that if mask wearing in the population reaches 80 %, it is expected that
the spread of SARS-CoV-2 will eventually decrease. Most importantly, when social isolation and mask wearing are implemented at the same time, as long as the population mask coverage rate reaches 30%, the goal of ultimately eliminating the spread of the virus in the community can be achieved.20

Another tracking analysis of nearly 200,000 people in different communities in the United States showed that the incidence of COVID-19 in communities that maintain good social distance is nearly 31% lower than that in communities that do not adopt social distancing. In some communities with high population density and relatively poor sanitation conditions, if people can spontaneously wear masks, the incidence of COVID-19 is expected to be reduced by 63%.21

Although there are disadvantages such as high population density and relatively weak health infrastructure, we found that this epidemic was quickly controlled in some countries, especially mainland China, because of effective government intervention with the active cooperation of the public, including social quarantine and mandatory wearing of masks in public, which was implemented quickly and compulsorily during the high-incidence period. In Europe and the United States, where the level of medical care and health is relatively developed, the government can only recommend but not enforce the above methods. As a result, the number of confirmed cases has remained high.

■ Conclusions
In summary, as a new type of emerging infectious disease, priority response is controlling transmission of COVID-19 first. In-depth research on pathogenesis, epidemiological characteristics, treatment methods and vaccine use will ultimately control viruses and eradicate diseases. At this stage, nonmedical interventions, especially wearing masks and social distancing, are effective and necessary methods. Moreover, the cooperation and conscious implementation of these methods by everyone in society is needed to overcome the pandemic.

■ Acknowledgements
I would like to thank my mother for her guidance of writing this paper.

■ References


**Author**

Shuying Liu is a Grade 2 student studying at the Kunming No.1 High School, Yunnan, China. She developed her interest in biology when she was 12. She hopes she can work on the field of Medical Biology in the future.
DNA Origami: Its Evolution and Application in Nanotechnology

Kavi A. Ullal
UWC-USA High School, State Rte. 65, Montezuma, New Mexico, 87731, USA; kavi.ullal@gmail.com

ABSTRACT: DNA origami is the construction of DNA based nanostructures, either through folding a single-stranded DNA scaffold or by creating tile-based structures. This paper reviews the conceptual advances and seminal research in this emerging field, and its application in areas like medicine and biotechnology. DNA origami has its roots in the discovery of the Holliday junction, a temporary branched nanostructure formed during meiosis. Nadrian Seeman pioneered DNA nanotechnology by using immobile Holliday junctions along with sticky ends and double-crossover motifs to create elementary structures. In the early 2000s, scientists transitioned towards using single-stranded DNA as scaffolds folded by several shorter oligonucleotides. Subsequently, researchers used shorter single-strands to create tile-based structures, equivalent to pixels in a molecular canvas. Binding between tiles on the canvas was selectively prevented through the use of edge protectors to arrive at a desired shape. More recent techniques such as fractal assembly have combined elements from the single strand scaffold and tile-based methods. DNA origami design has been further revolutionized by the development of software packages such as CaDNAno and Adenita. These packages will accelerate the application of DNA origami in nanofabrication and drug delivery, making it one of the most exciting and innovative areas of nanotechnology.

KEYWORDS: Biochemistry; Structural Biochemistry; Nanotechnology; DNA Origami; Holliday Junctions; DNA Scaffold.

Introduction

Scientists have long dreamt of being able to control the composition and structure of materials on the nanoscale. In the past two decades, DNA has emerged as an important building block for creating nanostructures. Scientists first considered DNA for this purpose due to its 2 nm diameter, structural repetition arising from its 3.37 nm helical pitch, and mechanical stiffness (with a persistence length of around 50 nm). These physical characteristics in combination with the programmable nature of its molecular makeup make it appealing for nanoscale construction. Many DNA based structures already occur naturally during cell division and other biological processes and can serve as a mechanistic basis for the artificial synthesis of nanomaterials.

DNA origami is the construction of DNA based 2D and 3D nanostructures, typically by folding a long single stranded DNA (scaffold) or using it to create tile-based structures. This design process has been transformed by the advent of new software packages such as CaDNAno and Adenita. These software tools are likely to make 3D DNA nanofabrication accessible to scientists in a wide variety of fields.

Recent innovations in DNA nanotechnology have exciting potential application in fields such as microbiology, photonics, and medicine. Although research regarding the use of DNA origami commercially is in the early stages, it has shown itself to be promising in the lab setting. DNA origami has been used to deliver drugs to cancer tumors in mice with high specificity. Scientists have also experimented with the construction of a controllable, 3D DNA origami box for application in drug delivery systems. DNA is also being proposed as a medium for nanocomputing, and early experiments have shown that it can do so. DNA origami appears likely to be a growing field for innovative research and has developed significantly since the original breakthroughs in DNA nanofabrication in the 1980s.

This paper is based on a literature review by the author of key publications in the field of DNA origami and nanofabrication. It summarizes the key intellectual advances and applications in the field. Existing literature reviews on DNA origami either focus on specific sub-topics or are written for specialized advanced researchers. This paper is written for high school or college students with an advanced interest in structural biochemistry, and for scientists who work in other fields who want an introductory primer on DNA origami. Making the advances in DNA origami more accessible to a wider audience will broaden the interest and possibilities in this fast-growing domain.

Methods

The Structure of DNA:

There are two approaches to forming nanoscale structures: top-down and bottom-up construction. Top-down construction is when the amount of material used for patterning and feature size is reduced to the required dimensions to create the desired shape. This differs from bottom-up construction, where structures are created by controlling the assembling atoms in the desired locations of interest. DNA has emerged as an important component of bottom-up nanofabrication techniques.

The unique nature of DNA makes it highly suitable for use in the construction of nanoscale structures. DNA consists of two strands intertwined and running in anti-parallel orientations. The strands are connected through pairs of nitrogenous bases. The four bases are Adenine, Thymine, Guanine and Cytosine. Adenine and Thymine always bond with each other through two hydrogen bonds, while Guanine and Cytosine pair through three bonds, resulting in a higher melting point. The
process of forming the double stranded molecule through base pairing is referred to as hybridization. Another feature of DNA is the ability to connect two DNA molecules by using strands that are of slightly different lengths, a process known as sticky end cohesion (Figure 1). If the ‘overhangs’ from two different helices are complementary, they can be induced to bind, thereby enabling the creation of larger lattice structures. The highly programmable nature of DNA means that it can be used for creating origami nanostructures.

**Holliday Junctions as Nanotechnology Building Blocks:**

The origins of bottom-up DNA origami can be traced back to the discovery of the Holliday junction. A Holliday junction is a naturally occurring temporary DNA nanostructure that forms when aligned strands of DNA break and cross over one another, prior to cell division. Holliday observed that four single DNA strands are centered around a branch point to form a junction-like shape with four arms. (Figure 2)

In the early 1980s, Nadrian Seeman, regarded as a pioneer in the field of DNA nanotechnology, began using the concept of the Holliday junction to create elementary designer nanostructures. Seeman recognized that naturally occurring mobile Holliday junctions have strands that are symmetrical, which can lead to the issue of branch migration. Branch migration is the tendency of the branch point of the junction to move along the branch strand. This would lead to insufficient rigidity when creating larger structures. To create rigid 2D and 3D nanostructures Seeman needed to convert Holliday junctions from being transient to having a more permanent shape. He proposed the immobile Holliday junction, which relied on selecting strands that were asymmetric about the junction. The asymmetrical junctions were artificially created to lock the DNA strands in position, thereby preventing branch migration.

To combine immobile Holliday junctions to form nanostructures, Seeman used sticky ends. As discussed earlier, when a portion of the end of a double-stranded DNA helix strand is left as an uncombined single strand (known as an overhang), it will have an affinity for the complementary base pair sequence of another single stranded DNA. When a certain Holliday junction has an overhang, it can bind to another Holliday junction with the complementary base pair sequence. Applying this method to several Holliday junctions facilitated the construction of basic nanoscale tiles.

However, one of the problems of using DNA branched motifs held together by sticky ends was a lack of sufficient mechanical rigidity to create periodic arrays and nanostructures. This was a significant stumbling block in the late 1980s and early 1990s for scientists in the field. A breakthrough was made when it was recognized that junctions could be reinforced with the addition of further helices. This involved making a hybrid of an immobile Holliday junction (J) and a double-crossover (DX) molecule. Double crossovers can form where two helical domains make contact, known as the crossover point. (Figure 3) This DX+J hybrid structure provided the necessary stiffness required to create robust branched DNA motifs. Seeman later stated that the discovery of the DX molecule as a rigid motif was a very significant leap forward in the field of DNA nanotechnology.

The DX+J hybrid structure was used by Seeman in the self-assembly of DNA to create 2D arrays and crystals. Seeman later also constructed the first 3D closed polyhedral object, a cube-like structure with six faces and eight vertices that corresponded to the branch points of junctions.

Scientists subsequently began to experiment with a range of motifs and tested their rigidity. This included paranemic crossover (PX) motifs as another means of combining DNA. Paranemic crossovers are like double crossovers in that they too form at the crossover point between two helices. However, while double crossovers form only at certain crossover points, paranemic crossovers form at every possible point of contact. With such structural motifs available, scientists have many options when it comes to the combination of DNA.
Single-Stranded DNA Origami:

In the 2000s, scientists took an alternative approach to creating DNA based nanostructures. They built on Seeman’s approach of creating structures that form by the process of molecular self-assembly, relying on the specific bonding between DNA base pairs. However, rather than using Holliday junctions and crossover-based hybrid motifs as building blocks, they used single-stranded DNA as the basis for the construction of DNA structures.

The concept of using single-stranded DNA folded into specific shapes using oligonucleotides was pioneered by William Shih and co-workers in 2004.¹⁸ In their method, a 1.7k nucleotide single DNA strand was folded into an octahedron through hybridization with 5 short oligonucleotides. The octahedron has 12 edges, or struts, joined at 6 vertices. The folding of the long scaffold strand was designed to take place in two stages. The first step is when the scaffold and the five associated oligonucleotides assemble to form a “branched-tree” structure. Here, double crossovers between the scaffold and the oligonucleotides form 5 of the 12 struts in the structure. In the second step, the remaining 7 struts are formed as paranemic crossovers, creating the octahedron shape. According to Shih, these two steps are interchangeable, and the order will not affect the formation of the octahedron. Shih’s method can be considered the first of its kind in the area of single-stranded DNA origami.

Just two years later, Paul Rothemund introduced his scaffold method,¹⁹ in which a single DNA strand is taken and folded into the desired shape using crossovers facilitated by several smaller oligonucleotides. Rothemund’s method differs from Shih’s as he uses a single, long 7.2 kilobase M13mp18 scaffold strand. This strand is a naturally occurring bacteriophage, whereas Shih’s main scaffold was artificially created. The base pair sequence of the M13mp18 strand is known and fixed, meaning that Rothemund’s method could be more easily replicated.

To fold the scaffold strand, Rothemund employed the use of 200 shorter oligonucleotides, or staple strands, as he called them. The design process begins by selecting several double-helices and shortening them sequentially to fit the constraints of the shape being designed, as shown below. To maintain the structure of the double-helices, several crossovers are used between them. Next, the scaffold strand is routed through the double helices such that it, at any point, is part of one of the two strands of the cylinder.²⁰ (Figure 4)

In order to create the DNA origami structures, the scaffold and staple strands are folded using one-pot annealing i.e., the mixture is rapidly heated to near boiling point, and then allowed to cool slowly to allow for self-assembly. The length of this annealing process is dependent on the complexity of the desired DNA origami shape. The structures are then purified by separating out properly formed structures from other molecules, based on their molecular weight. This is typically done using a technique known as agarose gel electrophoresis. The resultant nanostructures are then observed using a microscope with sufficient resolution, such as transmission electron, atomic force and super-resolution optical microscopes. Upon mixing, heating and cooling the mixture of scaffold and staple strands, well-formed structures with yields greater than 90% can be observed under a microscope. Rothemund used his scaffold method to create highly well-known shapes such as a smiley face, and a map of North America.¹⁹ Rothemund’s scaffold method truly made DNA origami more accessible to the scientific world.

Tile-Based DNA Origami:

While Rothemund’s scaffold method is easily replicable as it involves folding a specific DNA strand (M13mp18 bacteriophage), it does have its limitations. One limitation is that the size of the structures that can be created with it is fixed. The M13mp18 strand is 7,200 nucleotides in length, which corresponds to the construction of origami structures no more than 100 nanometers in diameter.²¹ This limited size of DNA origami structures is not an inherent feature of all DNA origami. It is strongly dependent on the shape and scaffold strand chosen. Additionally, the construction of a new shape requires a brand-new set of staple strands to be synthesized.

In 2012, Peng Yin and co-workers introduced the Single Strand Tile (SST) method, using a molecular canvas.²² While Nadrian Seeman’s original work with Holliday junctions could be considered a part of the tile method, modern methods to create tile-based structures take a different approach. In the SST tile approach, each tile or pixel is a unique, 42-base single-strand with four binding domains. In double stranded DNA, 42 bases correspond to four turns of the helix. During self-assembly, an SST tile can bind to four of its neighbors through hybridization, resulting in a lattice consisting of parallel DNA helices that are connected by single stranded linkages (Figures 5a and b). If a certain pixel is not desired to be a part of the final shape, binding is prevented by using edge protectors on its domains. The edge protectors have a sequence complementary to the uncombined binding domain, preventing further combination between certain pixels.

The rectangular lattice of SST tiles is equivalent to a molecular canvas consisting of pixels, each of which can either be included or left out of the final desired shape. The edge protectors are used selectively to exclude certain pixels by preventing them from binding with their neighbors. The strands are held at a temperature above the melting point of the DNA to allow all the DNA strands to separate, followed by a cooling protocol during which they self-assemble. The pixels that are not part of the intended design drop out of the

**Figure 4**: The process of designing a shape using the scaffold method. The double-helices are cut sequentially in order to fit into the constraints of the desired shape. Periodic crossovers are used to maintain the structure of the shape. This image is the author’s helical adaptation of an image from Paul Rothemund.
structure, leaving a 2D rendition of the chosen shape. (Figure 5c) The single-stranded tile method allows for a wide variety of shapes to be created. This rationale has been extended by Yin and co-workers into 3D, where they have made shapes as complex as teddy bears and hourglass figures.²³

The step prior to fractal assembly is synthesis of the individual squares. Each square is formed by a circular scaffold strand as well as edge, interior and bridge staple strands. To create an \( n \times n \) array requires \( n^2 \) test tubes of distinct tiles. In the first stage, test tubes are mixed in groups of four, resulting in the formation of \( 2 \times 2 \) arrays. At each subsequent step, a larger square array is created from four test tubes of smaller square arrays, that are themselves the product of the immediately preceding stage. The final product thus takes \( \log_2 n \) steps.

At any given stage it is necessary to ensure that the corresponding heating cycle does not cause the disassembly of previously created tiles. Qian and co-workers designed the binding edges such that edges that bind in the first step involve the highest number of binding domains and therefore the highest melting points. The melting points of the binding edges and therefore the temperatures required for assembly are designed to systematically decrease with each stage (Figure 6 right panel). This ensures that the binding that occurred at higher temperatures is not reversed. Using fractal assembly, the authors were able to create realistic depictions of complex images, such as that of a rooster and even the Mona Lisa.
specific base pair sequence to be able to bind to select points on the scaffold. One the scaffold strands has been selected and the staple strands have been synthesized, a solution of the two that has been annealed should result in structures that can be observed under a microscope.

While CaDNAno has a simple, intuitive interface and is the most widely used software for DNA origami construction, it has a few limitations. The structures that can be designed using CaDNAno are limited to parallel double-helices structures on either a honeycomb or square lattice. Structures such as 3D tiles, for example, cannot be designed on CaDNAno as they cannot be constrained to this lattice. As the size and complexity of DNA nanotechnology grows, the software that is used to design the structures needs to match it.

To address this, de Llano and co-workers introduced Adenita in 2020, a software tool to design 3D structures including but not limited to the ones that can be designed with CaDNAno. Adenita has the capacity to import designs from CaDNAno, while also creating other structures such as DNA tiles and wireframe structures. This is possible using the Daedalus algorithm, which can convert a 3D DNA origami object into the DNA sequences that are required for its self-assembly. Using the derived sequences, the desired structures can be practically synthesized. Unlike the other software of its kind, Adenita also allows the use of molecules other than DNA (such as proteins) as the building block for the construction of nanoscale objects.

The introduction of software over the past few years such as CaDNAno and Adenita has made DNA origami more accessible to scientists across the field. These tools have streamlined the design process, and the user interface is simple enough that even a high-school student can design basic structures using them!

**Applications of DNA Origami:**

In its infancy, DNA origami attracted attention due to the creation of unusual shapes or objects using bottom-up self-assembly. However, the field has since progressed well beyond being a mere curiosity or art form. The highly programmable nature of DNA means that it can be used to construct versatile nanoscale structures with practical applications. In the next few years, these structures can have immense use in nanofabrication, medicine, and nano-computation.

In the field of medicine, DNA origami can be used to carry and transport certain molecules, drugs, or proteins. These materials are generally carried on staple strands and given that there can be over 200 staple strands in a standard shape, there is plenty of area for them to be positioned. Zhang and coworkers used this method to deliver doxorubicin, a drug that can treat cancer, to target tumors in mice. Using the M13mp18 scaffold strand along with several staple strands, the authors constructed a DNA origami triangle. The triangle was incubated with doxorubicin for 24 hours, and, following the drug delivery, a fluorescence imaging technique was used to identify the size of the tumor. The mice in which the drug was delivered using DNA origami were found to have tumors significantly smaller than those without the DNA origami delivery over a 12-day period. The primary reason for this is that the double helices that make up the triangle have a significantly large number of binding sites, meaning the drug can be delivered with great specificity to the site of the tumor. Most conventional methods of drug delivery lack this specificity. This experiment demonstrated that DNA origami can have great use in nanofabrication and healthcare.

While several 3D DNA origami structures have been created, from prisms and tetrahedrons to complex octahedrons, prior to 2009, none of these structures had a fully closed surface along with a hollow cavity capable of carrying molecules or drugs. Kjems and co-workers were able to achieve this when they constructed a hollow, 36 x 36 x 42 nm³ 3D DNA origami box from the M13mp18 DNA single strand. This box could also carry a cargo load and was controllable in that it could open on the insertion of two keys into their respective locks located on the exterior of the box. The locks had specific, 8 nucleotide sequences that corresponded to particular DNA keys. When inserted, the box would open if the sequences matched. This DNA origami box was used to carry and deliver drugs to cancer cells.

However, this novel design lacked the ability to repeatedly open and close. In 2012 the Kjems group introduced another iteration of their 3D DNA origami box. Here, the box was switchable, meaning it could open and close several times, and it was 1/7th the size of the original box.

Given the highly controllable manner in which DNA can be folded into desired structures, DNA origami has also been experimented with as a way to inhibit viruses. This is done through the targeted binding of the DNA origami structural motifs to the exterior proteins present on viruses. Wang and co-workers began experimenting with this method in early 2020, when they designed a star-shaped DNA origami structure through tile self-assembly. The ten vertices of the star were formed by a hair loop-like bend of a single strand. This gave the star the structural flexibility necessary to bind to the desired virus under various physical conditions. The vertices of the star also served as the location for ED3-targeting aptamers (single strands capable of folding into specific shapes to bind to targets). The ED3 structure is found on the surface of the dengue (DENV) virus, which the star was designed to inhibit. This DNA-based construction was able to bind to the ED3 structure with high specificity.

The DNA origami star was designed such that when binding between the star and virus took place, it would cause a
change in the external environment of dye molecules attached to it. This would result in a fluorescence signal that could easily be detected, indicating that binding between the star and virus was complete. Wang and co-workers’ method has begun to be used commercially by Atom Bioworks29, a firm that maps the arrangement of surface proteins of viruses for use in therapeutics and medicine. In light of the global medical and economic consequences of COVID-19, the application of DNA origami to detecting and inhibiting viruses is likely to attract significant research and commercial interest.³⁰

Another area in which DNA origami has seen commercial development is optics and microscopy. Gattaquant³¹, a biotechnology company based in Germany, has been one of the first companies to begin using DNA origami for commercial use. Gattaquant uses DNA origami to create several nanoscale structures, from nanorulers to molecular breadboards. The company offers customers these structures for use in super-resolution microscopy, which helps in positioning molecules in desired locations and orientations. Gattaquant’s nanorulers can be used to check the resolution of images, and have the ability to serve as a tool to align microscopes.

Discussion and Conclusion

As seen in this review, DNA origami has progressed significantly since its roots in early DNA nanofabrication done in the 1980s. Nadrian Seeman provided the early significant breakthroughs, by recognizing that immobile Holliday junctions could form a basic structural unit, particularly when its rigidity was augmented by double and paramenic crossovers. An important advance in the early part of this century was the use of single stranded DNA scaffolds in combination with smaller staple strands to design nanostructures. More advanced tile-based methods then emerged that addressed some of the drawbacks of using a long single strand scaffold. Recent research work has focused on combining elements of the scaffold and tile-based approaches. All these design efforts have been significantly transformed with the launch of software design packages for DNA based 3D nanostructure modeling. In aggregate, these advances represent rapid scientific progress in an exciting field with significant real-world applications.

Looking ahead, research in DNA origami and nanofabrication is likely to develop in two broad directions. The first is the refinement of DNA origami techniques themselves, to find methods with higher replicability, cost efficiency and yields. The second is research that can support the transition of DNA origami to detecting and inhibiting viruses is likely to attract significant research and commercial interest.³⁰

References

Author
Kavi Ullal is a student at UWC-USA High School. He has always been passionate about biochemistry and would like to pursue it in his later career. Kavi has been involved in projects in his local community to solve issues pertaining to the availability of natural resources.
Familiarity-Dependent Relaxation Responses of Classical Musicians to Music Selections

Veronica Li
Princeton Day School, 650 Great Rd, Princeton, NJ 08540, USA; veronicaliclarinet@gmail.com

ABSTRACT: Performance anxiety is a prevalent mental health issue among professional musicians. Pre-performance music therapy including a listening component has been shown to help musicians relax. While substantial research has explored which compositional features of classical pieces induce relaxation, this study hypothesized that for professional musicians, familiarity with a relaxing classical piece as well as performance attributes may impact relaxation level. Familiar pieces may produce a lower relaxation level when the musician's performance expectations are not met, while unfamiliar pieces — bearing no such expectations — might more readily induce relaxation. Musician and non-musician participants were recruited to listen to a familiar classical piece performed in two markedly different styles, and also to a generally unfamiliar piece. The subjects evaluated each piece, rated the relaxation level experienced, and provided comments. For musicians, the unfamiliar piece generated significantly greater relaxation levels than the familiar piece (Familiar mean = 7.1, SEM = 0.29; Unfamiliar mean = 8.2, SEM = 0.42; p = 0.026). Furthermore, the less that the familiar piece’s performance met a musician’s expectations, the less relaxing they found it (R = 0.54, p = 0.00029). These findings may help determine which classical pieces best alleviate professional musicians’ performance anxiety.

KEYWORDS: Behavioral and Social Sciences; Clinical and Developmental Psychology; Classical Music; Music Therapy; Classical Musicians; Relaxation.

Introduction

Mental health and stress are common obstacles to happiness for many Americans. According to the National Institute of Mental Health, almost 1 in 5 US adults are living with a mental illness of varying categories and levels of severity. The recent public struggles with mental health by elite athletes Simone Biles and Naomi Osaka have raised the issue of performance-based mental health concerns within the public consciousness. Research into the alleviation of stress caused by performance anxiety may therefore prove beneficial to those coping with similar struggles.

Excessive stress has a negative impact on human physical and psychological health, and musicians are no exception. A widely quoted survey initiated and funded by The International Conference of Symphony and Opera Musicians in 1986 raised researchers’ awareness about the impact of stress upon professional musicians. This survey had more than 2,000 respondents, with 24% of musicians reporting they experienced stage fright, 13% reporting experiences of acute anxiety, and 17% experiencing depression. Since then, studies have addressed multiple components of musicians’ performance anxiety (MPA). Researchers have explored the physical manifestations of performance anxiety and the factors that contribute to it, exploring such topics as variations in heart rate and blood pressure varying degrees of anxiety before, during, and after a performance; higher levels of the stress hormones cortisol and cortisone in singers performing in public compared to singing without an audience; the relation between a musician’s attitude before a performance (perceived as either a threat or a challenge); and the degree of success of the performance.

In a parallel line of research, music therapists have investigated the impact of interventions with relaxing music to induce relaxation for general populations, including high-anxiety populations, such as parents of children receiving hospital treatment. Establishing the criteria for the compilation of lists of music pieces that induce relaxation in listeners has subsequently become an important line of research. Such studies have focused on determining structural features that bestow a “relaxing” quality on a piece of classical music (defined here as notated music composed between 1700 and 1900 meant for performance in concert settings). One of the most thorough studies, which explores different common musical properties found in relaxing music pieces, is “The Interplay of Preference, Familiarity and Psychophysical Properties in Defining Relaxation Music.” This research has concluded that the most relaxing pieces are those that are familiar, slow, soft, with minimal changes. The author even determines specific descriptions and values for relaxing properties (e.g., the melodic contour must be homogeneous, gentle, or the tempo must be between 60 to 80 bpm).

Building on the successes of music therapy for the general population, music therapists developed complex pre-performance therapy sessions for professional musicians with high levels of MPA. Therapeutic techniques used include listening to classical music along with breathing and muscle relaxation exercises. In view of the fact that musicians’ relationship with music is different than the one experienced by non-musicians, it seems likely that the music they may find relaxing may not always be the same as that for general audiences. For example, Tan et al. found there was a significant positive correlation between familiarity of a piece
and its relaxation effect for non-musicians, while this was not the case with the music therapists participating in their study. They proposed this might be due to the therapists associating the music with their work and thus finding it less relaxing. A similar effect might be expected for classical musicians, for whom playing familiar music is their work. However, Tan et al. note that no previous studies have explicitly explored the relationship between professional musicians’ relaxation and familiarity. Further, to the best of our knowledge, no study has looked into whether the ability of a particular performance of a familiar piece to induce relaxation is affected by whether it meets the musician’s prior, personal expectations of it.

The objective of this study based on music-assisted surveys is to obtain information narrowly focused on professional, classically trained musicians’ reactions to pieces that are considered to be relaxing (according to Tan et al. criteria) in order to tailor the lists of classical music compositions used in music therapy sessions for such musicians.

The present survey tests the hypothesis that the relaxation level experienced by a classically trained musician while listening to a familiar piece may depend as much on how its particular performance conforms to a musician’s expectations as on its composition features. Musicians’ expectations regarding the performance of a given piece may be higher when it is familiar. They may harbor strong opinions about the “correct” tempo, dynamics, phrasing, instrumentation, etc. for a piece with which they are familiar. Consequently, listening to an interpretation that does not fulfill these expectations may interfere with feelings of relaxation. Conversely, musicians may have reduced expectations regarding the interpretation of an unknown or lesser-known piece that conforms to the criteria established for “relaxing music,” so such a piece may have a greater likelihood of inducing a higher degree of relaxation.

This survey contributes to the field of music therapy by addressing the need for guiding principles in the selection of classical music pieces meant specifically for the relaxation of professional, classically trained musicians. It may recommend narrowing the listening selections meant for musicians whose preferences are not known by the therapist to pieces that are less likely to be familiar, while still possessing the general qualities known to induce relaxation.

**Methods**

**Participants:**

The study recruited two groups of respondents with IRB approval. All participants were aged 30 and up and recruited largely from New York and New Jersey, where this study was based. The first group consisted of classically trained professional musicians (mostly area teachers and performers). The second group was a control group of non-musicians (defined as persons who have not had formal music instruction, nor participated in band, orchestra, or choir past middle school).

Potential participants were approached by email or verbally with an invitation to participate in this survey. They were asked to provide consent in writing or verbally.

Of 46 prospective participants, 20 musicians and 18 non-musicians returned surveys for at least two of the pieces and were included in the study analysis. The demographics of these two groups were largely similar in terms of gender (60% female musicians and 76% female non-musicians) and age (median age range was 50-59 for both groups). There was some difference in distribution across race-ethnicity, with the musicians being majority Caucasian (85%), while non-musicians were a more diverse group with the largest groups being Asian (43.75%) and Caucasian (37.5%).

**Choice of Pieces:**

All respondents were asked to listen to the same three orchestral pieces. The first two pieces were chosen for their general familiarity and the third for its general unfamiliarity. All pieces were selected in conformity to the principles laid down by Tan et al. (i.e., slow tempo, no abrupt changes in dynamics or rhythm, gentle melodic contour). Care has been taken that these pieces be performed by reputable ensembles. The titles and performers were not made known to the participants.

Piece #1, Johann Sebastian Bach’s *Air from Orchestra Suite #3*, BWV 1068, is a widely known composition that appears on the majority of compiled lists of relaxing pieces.³² It is performed by an early music ensemble with period instruments. It will henceforth be referred to as “Familiar 1.”

Piece #2 is the same piece as the above but performed in a more modern interpretation by a symphonic orchestra with a harpsichord. Thus, it is more likely to conform to modern musicians’ expectations of this piece. It will henceforth be referred to as “Familiar 2.”

Piece #3, Ignatius Pleyel’s Andante grazioso from the Symphony in F Minor, Ben. 138, is a classical piece that is not frequently performed and not studied in music history classes, therefore likely to be unknown to most participants. It will henceforth be referred to as “Unfamiliar.”

**Survey Questions:**

Participants were emailed a cover letter including the objective, detailed instructions, a consent form, and links to three surveys. For each one, participants were asked to listen to a piece of classical music and answer questions about how relaxing the piece felt to them. Optional demographic questions appeared at the beginning of the first survey. Each survey took approximately 15 minutes to complete. Each respondent completed all three within a week, at least one day apart from each other at about the same time of day.

The surveys are available for review at the following links as of August 2021.

Survey 1: https://083qvhq080.typeform.com/to/oPqR1f7M
Survey 2: https://083qvhq080.typeform.com/to/e5FqwJX6
Survey 3: https://083qvhq080.typeform.com/to/kBjrVtKf

Two of the major metrics for this study concern level of relaxation and types of emotion evoked by the music listened to. The scale used to measure the participant’s relaxation is commonly used in music therapy studies.¹⁰ The list of emotions possibly elicited by the music comes from the widely accepted “Thayer’s model.”
Results and Discussion

Validation of Relaxation and Familiarity:

As expected, all three pieces had a positive relaxation effect on the majority of participants (27/35 for Familiar 1, 24/35 for Familiar 2, and 25/33 for Unfamiliar). This was true for both musicians and non-musicians (Figure 1).

Figure 1: All three pieces most commonly had a positive effect on both musicians and non-musicians. However, the non-musicians had a greater positive response to the familiar pieces than the unfamiliar piece while the musicians had a greater positive response to the unfamiliar piece. ("Primarily Positive": selected words included at least one of happy, pleased, relaxed, peaceful, and calm. "Primarily Negative": selected words included at least one of annoying, angry, nervous, and bored. Mixed: selected words included at least one positive and one negative term. "Other": the options included exclusively excited, sad, sleepy, and none of the above.)

Notably, however, proportionally more non-musicians found the Familiar pieces to be relaxing, while proportionally more musicians found the Unfamiliar piece relaxing. This difference will be explored more quantitatively in the next section.

100% (20/20) of musicians and 83.3% (15/18) of non-musicians confirmed that they were familiar with the well-known baroque piece presented in two different interpretations (Familiar 1 and Familiar 2). In contrast, the less well-known classical piece (Unfamiliar) was known to only 5.3% (1/19) of musicians and 6.25% (1/16) of the non-musicians. Because the central hypothesis of this study is that familiar music is less relaxing to musicians than unfamiliar music, in all subsequent analysis, the small number (8/111) of responses where there was a mismatch between a respondent's actual familiarity with a piece and the intended familiarity have been excluded (i.e., responses about the Familiar pieces from respondents who were unfamiliar with them and about the Unfamiliar pieces from respondents who were familiar with them).

Musicians Found Familiar Pieces Less Relaxing:

The hypothesis of this study proposed that musicians might be more relaxed when they listen to unfamiliar music as opposed to familiar music due to their expertise in music and inclination to critique the music they hear against their expectations of it. In order to test whether or not the familiarity of music did in fact have an effect on musicians' relaxation responses, we asked all participants to score their relaxation response to the Unfamiliar and Familiar pieces on a scale of 0 to 10. As hypothesized, the musicians ranked the Unfamiliar piece as significantly more relaxing (Figure 2: mean Familiar = 7.1, SEM = 0.29; mean Unfamiliar = 8.2, SEM = 0.42; p = 0.026), while the non-musicians did not have a statistically significant difference in their relaxation ratings for the Unfamiliar versus Familiar pieces (mean Familiar = 7.9, SEM = 0.38; Unfamiliar = 8.1, SEM = 0.44; p = 0.75).

Figure 2: Both musicians and non-musicians gave their relaxation response to familiar and unfamiliar music on a scale of 0 to 10. Musicians' relaxation levels were significantly higher for the unfamiliar piece compared to the familiar piece, while the relaxation levels for the non-musicians were not. (Student’s T-Test, p<0.05. Error bars represent 2 SEMs)

Musicians Engaged More Critically With Familiar Pieces:

To explore whether the musicians were more critical of familiar than unfamiliar music, surveys asked musicians to select positive musical characteristics (from a choice of 6) they agreed with and did not agree with for each of the three pieces after listening to each piece.

Remarkably, for the Unfamiliar piece, almost half (8 out of 17, 47.1%, SEp = 24%) agreed with all the positive aspects of the performance and only 3 participants (18%, SEp = 19%) disagreed with at least one of them (Figure 3, note that these percentages do not add up to 100% because participants were allowed to neither disagree nor agree). In contrast, for the Familiar pieces, these numbers were essentially flipped, with few musicians agreeing with all positive aspects and many disagreeing. Only 6 out of 40 (15%, SEp =11%) responses agreed with all aspects of the performance for the Familiar pieces, while more than half disagreed (27 participants, 67.5%, SEp =15%).

 Across all three pieces, musicians most commonly criticized dynamics. Of the musician survey responses for this question, 14/44 responses disagreed with the Familiar pieces’ dynamics. For the Unfamiliar piece, 2/18 people disagreed with dynamics.

These results demonstrate that professional, classically trained musicians are more likely to be lenient in judging the performance of a piece they do not know than a piece that they do. This finding that musicians were less critical of the Unfamiliar piece is interesting given the previous findings that musicians had more relaxed feelings and a higher relaxation level associated with the Unfamiliar piece (Figures 1 and 2 respectively). The less critical approach to less familiar music may explain why these musicians found the unfamiliar piece more relaxing. It furthermore raises the question of whether when a familiar piece does meet a musician’s expectations it can be as relaxing as an Unfamiliar piece despite being subjected to higher expectations. This possibility is addressed in the next section.
that who indicated level 10 of relaxation for both Familiar 1 and Unfamiliar piece, it also conformed to the listener’s particular when a familiar piece was experienced to be as relaxing as the piece relaxing if it meets their performance expectations. In of relaxation). This demonstrates that musicians may find a piece (with the other Familiar piece having the lowest level of relaxation) for one of the two performances of the same level of relaxation for one of the two performances of the familiar piece (Familiar 1 or Familiar 2) as for the Unfamiliar piece. Several participants indicated that the act of analyzing classical music can make it difficult to relax. One participant stated, “It is also hard to truly relax when you’re critiquing the performance.” Another explained, “As an experienced classical musician I find it hard to relax to classical music. I find myself analyzing it, trying to figure out who might have written it, what the form is, how it might have been a more effective performance.” Furthermore, they suggested that within the broader category of familiar music, musicians found their relaxation was particularly reduced if the piece did not conform to their particular expectations for it. Because the Familiar 1 piece is a marginal style, for most musicians it conforms less to their expectations than the Familiar 2 piece. As such, one participant observed about the Familiar 1 piece, “The phrasing felt choppy and the articulations a bit strange. It negatively affected my relaxation and made me feel a bit anxious.” Likewise, another participant commented on this same piece about the trade-off between analysis and relaxation: “The interpretation was not very familiar but after listening a bit (not relaxing but analyzing) it felt alright.” In contrast, this same participant wrote about the Unfamiliar piece, “Because I did not know this particular piece and also this period of musical style is not my main focus at work, I didn’t find any strong interpretational issues that would bother me (i.e., intonation, bad recording quality, exaggerate tempi) and I could just enjoy listening to it.” Overall, while the Unfamiliar piece was found to be the most relaxing, nevertheless 8/20 musicians actually indicated the same level of relaxation for one of the two performances of the familiar piece (Familiar 1 or Familiar 2) as for the Unfamiliar piece (with the other Familiar piece having the lowest level of relaxation). This demonstrates that musicians may find a piece relaxing if it meets their performance expectations. In each of these cases, the open-ended comments suggested that when a familiar piece was experienced to be as relaxing as the Unfamiliar piece, it also conformed to the listener’s particular critical expectations of that piece. For example, a musician who indicated level 10 of relaxation for both Familiar 1 and Unfamiliar pieces stated for Familiar 1: “The performance … met the stylistic performance practice of the Baroque period. The interchange of inner voice’ melodic lines and motifs was tastefully done. … Overall, it was a pleasant performance.” And about the Unfamiliar piece “very enjoyable performance … [technical details listed] … Excellent performance overall.” The musicians posit through these responses that for familiar pieces, a mismatch between their expectation and reality causes a decrease in their relaxation. Indeed, this possibility was borne out quantitatively as well. We found a negative correlation between the number of aspects of performance that a musician disagreed with for a given Familiar piece and the relaxation score they gave to it (Pearson’s Correlation, R = -0.5443, p =0.00029).

Taken together, these results demonstrate that when listening to Familiar pieces, participants were more inclined to critique the performance. When the performance didn’t conform with their expectations, this interfered with their relaxation. However, for the Unfamiliar piece, because it was something that was previously unknown to the musicians, they did not automatically have a need to critique it and were able to just listen to it and relax.

**Conclusion**

Music affects non-musicians and professional musicians differently. Professional musicians have a more trained ear in comparison to non-musicians and a more specific idea of what they look for in the performance of classical music pieces, especially when it comes to pieces they have heard multiple times. The hypothesis that musicians would be more relaxed when they listen to a piece that is unfamiliar to them rather than a piece that is familiar to them was tested.

In order to test this, participants listened to three pieces (Familiar 1, Familiar 2 and Unfamiliar) and rated their relaxation levels, as well as answered questions that could give insight about them, such as degree of familiarity and observance of positive musical characteristics.

Musicians on average had a higher relaxation response to the Unfamiliar piece than the Familiar pieces. This was in contrast to the non-musicians, whose relaxation response did not significantly differ for the Unfamiliar and Familiar pieces. This finding aligns with the supposition of Tan et al. that for people with professional musical training - in their case, music therapists - familiarity might not correspond to greater relaxation.¹⁰ However, in their study, Tan et al. found only a lack of positive correlation between relaxation and familiarity. On the other hand, in our study, the professional musicians’ relaxation level significantly decreased when they listened to familiar music. This strong effect could be due to the fact that musicians, who may actually have played these pieces, are likely to have a deeper familiarity with familiar pieces than music therapists.

These results showed that the professional musicians – many of them performers – are in general more critical of familiar pieces. This was supported by their comments that the act of trying to critique a familiar piece made it more difficult for them to relax. This strongly suggests that there is a causal relationship between their increased criticalness and
the decrease in relaxation when listening to familiar pieces. Comments that described how the particular performance style of Familiar 1 made it harder to relax because it was not what the participants expected to hear further suggested that the decrease in relaxation may be greatest when a familiar piece does not meet a musician’s expectations. This was supported quantitatively by a correlation between the number of ways in which a familiar piece failed to meet a musician’s expectations and the relaxation level they experienced listening to it. This shows that musicians are indeed able to relax while listening to familiar pieces, but on the condition that they meet their particular personal performance expectations.

These findings lend strong support to the hypothesis that familiar music is more likely to be less relaxing for musicians than unfamiliar music. Additional testing, involving a larger group of participants and a more diverse set of classical pieces, can be done in the future in order to provide a broader foundation for the generalization of these conclusions. It would also be interesting to explore whether classical music is most effective for relaxing classical musicians or if other genres that are not their working medium might be more effective.

Furthermore, researchers may be interested in detailing our observations by breaking down the familiarity aspect into two categories: familiarity acquired through listening and familiarity acquired through playing a piece and determine whether this difference in the degree of familiarity affects the relaxation level.

The conclusions of this study are relevant to the design of therapy sessions dedicated to professional classically trained musicians who experience performance anxiety. Given the strong connection of musicians to music, music therapy would seem to have the potential to be particularly powerful in helping them but could also require more nuance in how it is delivered. The results contribute to a better understanding of the mechanisms involved in musicians’ reactions to familiar versus unfamiliar music. Moreover, they provide a guiding principle in assembling lists of suitable classical music pieces for relaxation-oriented listening. In settings where the musician’s preferences are not known to the therapist, the safer approach is to have a selection of pieces that, while displaying the necessary features to make them relax, are less likely to be familiar. It is a narrowly focused yet meaningful contribution to the field of music therapy and through that to the wider field of mental health.

Acknowledgements
We would like to thank Dr. Tomasz Rzeczycki for helping us with the musical aspects of this paper. His musical expertise was critical in designing this experiment, such as the selection of pieces for participants to listen to. We would also like to thank the Princeton Day School REx (Research Experience) Program and the members of the Princeton Day School IRB Committee for all their support and for making this research possible.

References

Author
Veronica Li is a student at the Princeton Day School and a clarinetist at the Juilliard Pre-College. She is interested in combining her passions for science and music by exploring music therapy and music therapy research.
Rochester, New York
Hosted by Rochester Institute of Technology
@GeniusOlympiad
www.geniusolympiad.org

“Let’s build a better future together”

INTERNATIONAL ENVIRONMENT PROJECT FAIR FOR GRADES 9-12

STANDARDS
* Governance
* Education
* Culture
* Operations
* Quality

Lifetime Qualities
MORAL
Caring, Modest, Ethical, Respectful

THINKER
Knowledgeable, Inquirer, Curious

PRODUCTIVE
Achiever, Excellence Seeker

INNOVATIVE
Problem Solver, Explorer

VISIONARY
Global Citizen, Inclusive

ACASU BENEFITS
Enhances your institution’s reputation and garners the esteem of Americans.
Assures students and their families that your institution meets American standards for high quality education.
Shows students and parents your institution’s commitment to success and your ability to perform at high quality educational standards.
Shows your colleagues and the general public that your institution has a genuine interest in institutional improvement.
Sets your institution apart as a leader among others.
Provides you with a network of skilled professionals and resources to assist meeting your accreditation, enrollment, and college attainment goals.
Ensures that your standard operating procedures are compliant with U.S. international student regulations.

ACASU.ORG
is a publication of

N.Y. based 501.c.3 non-profit organization dedicated for improving K-16 education