SISE 2601

Projects and Datasets

Table of contents:

[Pre-approved projects using academic data 1](#_j10iihqj9s1n)

[ER Waiting times in Urgent Care units in Israel 1](#_o67xl3omdfcc)

[Engagement with AI-generated Health Information 2](#_afg6cpx5cp0t)

[Experiences of Older LGBTQ's in Israel 2](#_aqc02rrlifcq)

[Reward sensitivity in PTSD 2](#_xxo9rbtk6nrj)

[Calibrating Real-World Importance of Social Media 3](#_e1j761l3fi4u)

[Explore patterns in the crowdfunding platform Kiva 3](#_q3wctyxx1ajc)

[eQuality in Quality: Differences in scientific quality in Google Search 4](#_mvav976eqb2b)

[AI-generated Email Reply Suggestions 5](#_x7sjupfagh6l)

[Substance abuse: admission and discharge records 6](#_ypk9aijyrc50)

[Additional data sources 6](#_sgtd22c186hz)

## 

# Pre-approved projects using academic data

## ER Waiting times in Urgent Care units in Israel

עבודות ומחקרים רבים נעשו סביב שאלת העומסים בחדרי המיון ברחבי העולם. שיטות רבות פותחו ואומצו על מנת לייעל את תהליכי הקבלה, המיון, הטיפול והשחרור. הנחת היסוד של מחקרים אלו הינה שרמת עומס גבוהה במיון מובילה לאיכות טיפול ירודה, לעומס נפשי וגופני על הצוותים ולבזבוז משאבים. אחת התופעות אשר תורמות לשהייה ארוכה של מטופלים במיון מתייחסת לעיכובים של מטופלים על מיטות האמבולנסים אשר הביאו אותם למלר""ד (מחלקה לרפואה דחופה). תופעה אשר מכונה גם "עיכוב בפריקת אמבולנסים". עיכובים אלו הינם בעלי משמעות לזמן עד קבלת ובדיקת המטופל ועל כן גם משפיעים על הזמן עד תחילת הטיפול עצמו. בנוסף, לעיכוב אמבולנסים במיון ישנה השפעה כפולה על איכות הטיפול בקהילה, זאת כיוון שאמבולנס אשר לא מתפנה מהמיון אינו זמין לקבלת קריאות חירום נוספות וכך ישנה למעשה, פגיעה כפולה באיכות הטיפול בפרט ובזמינות שירותי רפואת חירום בכלל. התופעה גם משפיעה על עייפות פיזית ונפשית של צוותי האמבולנס וחושפים אותם לרמה גבוהה של אלימות מילולית ופיזית, כל אלו גורמים בסופו של דבר לנשירה מוגברת מהמקצוע. בישראל היקף התופעה והשלכותיה הבריאותיים והכלכליים אינם ידועים.

מטרתו העיקרית של מחקר זה לאפיין את היקף ופרטי התופעה של עיכוב מטופלים המובאים למלר״דים על ידי צוותי האמבולנס של אגודת מגן דוד -אדום בישראל. מטרות משנה יבחנו את הקשרים בין משך ההמתנה לבין מצבם הרפואי של המטופלים, משתנים תלויי זמן ומשתנים גיאוגרפיים.

**Where is the data?** Contact Prof. Odeya Cohen ([odeyac@bgu.ac.il](mailto:odeyac@bgu.ac.il)).

**Can multiple groups work on this?** Yes, but no more than TWO GROUPS with clearly distinguishable research questions and variables. Coordinate a time to meet with Prof. Cohen, then afterward write her a short email expressing your interest in the project and explaining why you should be selected.

**How to get more info about the data?** Contact Prof. Odeya Cohen ([odeyac@bgu.ac.il](mailto:odeyac@bgu.ac.il)).

## Engagement with AI-generated Health Information

How do people engage with everyday health-related information provided by GenAI?

A representative sample of 500 Hebrew-speaking Israelis filled a survey on June 2023 (early usage of GenAI). The data depicts their position regarding AI (attitudes, trust, deference for, perceived benefits and risks), their knowledge about AI and its epistemic limitations, and familiarity and use of GenAI technologies. In the survey, respondents were also asked to perform a task: to make a decision (open-ended questions) about an everyday health-related issue before and after they were introduced to information provided by GenAI. The dataset includes all respondents' answers as well as a complex content-analysis of their open-ended questions. We're interested in studying how people engage with the information GenAI provided and how it helped them in reaching the preferable decisions regarding everyday scientific issue. We are planning to have an additional data collection again around June-July 2024, to also look for changes over time.

Key research questions:

1. Who are the people to best engage with scientific information provided by GenAI?

2. What are the differences between use of and position toward GenAI (as a technology) vs. the information it provides?

3. What are the key changes in GenAI adoption in Israel a year later?

**Where is the data?** One SPSS file with 500 responses and 250-300 variables. Contact Dr. Inbal Klein-Avraham ([Inbal.kein@campus.technion.ac.il](mailto:Inbal.kein@campus.technion.ac.il)) for access.

**Can multiple groups work on this?** Yes, but no more than TWO GROUPS with clearly distinguishable research questions and variables. Coordinate a time to meet with the researcher, then follow up afterward with a short email expressing your interest in working on this project and explaining why you should be selected.

**How to get more info about the data?** Contact the researcher.

## Experiences of Older LGBTQ's in Israel

ד״ר קון-שוורץ אספה נתונים מכ-450 משיבים שמזדהים כלהט"בים בגילאי 50 ומעלה. המחקר התמקד בשאלות על הקשרים החברתיים שלהם, על הרווחה נפשית (דיכאון, חרדה, בדידות, שביעות רצון מהחיים), על חוויות של אפליה, הומופוביה מופנמת וגילנות. נתונים אלו מאפשרים לבחון מגוון של שאלות מחקר שונות לגבי האוכ׳ הנ״ל.

**Where is the data?** Contact Dr. Cohn-Schwartz ([ellasch@bgu.ac.il](mailto:ellasch@bgu.ac.il)) for access.

**Can multiple groups work on this?** Yes, but no more than TWO GROUPS with clearly distinguishable research questions and variables. Coordinate a time to meet with the researcher, then follow up afterward with a short email expressing your interest in working on this project and explaining why you should be selected.

**How to get more info about the data?** Contact the researcher.

## Reward sensitivity in PTSD

Analysis of a computerized task of people with different levels of post-traumatic stress disorder (PTSD) and its association with reward and punishment learning. The project will include generating a Bayesian reinforcement learning models of the learning behavior of participants in the study.

Research questions:

1. Does PTSD associated with impaired reward learning?

2. Which is the best RL model that will fit behavior of 4 bandit multiple choice task?

**Where is the data?** Contact Dr. Or Duek ([ord@bgu.ac.il](mailto:ord@bgu.ac.il)) for access.

**Can multiple groups work on this?** Yes, but no more than TWO GROUPS with clearly distinguishable research questions and variables. Coordinate a time to meet with the researcher, then follow up afterward with a short email expressing your interest in working on this project and explaining why you should be selected.

**How to get more info about the data?** Contact the researcher.

## Calibrating Real-World Importance of Social Media

Social media exhibit an immense variability in plausible effects on society — this presents a persistent challenge to social science. In addition to divergent findings suggesting beneficial (“good”) and harmful (“bad”) consequences, assumptions about effect sizes range from dramatic to insignificant, leading some researchers to warn against moral panics (Jungherr & Schroeder, 2021). We aim to benchmark social media’s capacity to interact with (offline) society as a whole. Such calibration is presented by linking a broad longitudinal sample of online activity with offline behavioral data. Using a sample of twelve years of Israeli online and offline behavior such as Covid-19 lockdowns, military campaigns, elections and holidays as natural experiments, we measure the strength of online-offline linkage in a series of multilevel linear models. The results will offer a robust empirical estimate of the plausible degree to which “good, bad, or ugly” social media affect reality (and vice versa).

To what extent do political (e.g., elections, wars), social (e.g., Eurovision, sports events), and religious (e.g., Saturdays, Jewish holidays) events affect the communication behavior on social media in Israel?

**Where is the data?** Contact Dr. Yossi David ([davidyos@bgu.ac.il](mailto:davidyos@bgu.ac.il)) for access.

**Can multiple groups work on this?** Yes, but no more than TWO GROUPS with clearly distinguishable research questions and variables. Coordinate a time to meet with the researcher, then follow up afterward with a short email expressing your interest in working on this project and explaining why you should be selected.

**How to get more info about the data?** Contact the researcher.

## Explore patterns in the crowdfunding platform Kiva

Kiva is a non-profit organization that operates a crowdfunding platform for providing microloans to individuals in need around the world, particularly in developing countries. Founded in 2005 by Matt Flannery and Jessica Jackley, Kiva aims to alleviate poverty and promote financial inclusion by connecting borrowers with lenders who contribute as little as $25 to fund a loan.

More about Kiva: Borrowers on Kiva's platform often use these microloans to start or grow small businesses, support their education, or improve their living conditions. Once a loan is funded, it is disbursed to borrowers through local microfinance institutions, called Field Partners, which are responsible for vetting, administering, and collecting repayments.

As borrowers repay their loans, the funds are returned to the lenders, who can then choose to fund other loans, donate the money to Kiva, or withdraw it. The organization has facilitated millions of loans and has a high repayment rate, making it an attractive option for individuals who want to make a positive impact on a global scale.

Possible data science questions

Here is a non-exhaustive list of questions you can tackle using Kiva's data to gain new insights and help improve the platform's effectiveness:

1. Loan distribution and impact: How are loans distributed across different countries, sectors, and genders? What is the impact of these loans on borrowers' socioeconomic conditions, and how does this vary by region and sector?

2. Loan repayment patterns: What factors influence the likelihood of timely loan repayment? Can we build a predictive model to identify loans with a high risk of default, based on borrower characteristics, loan amount, repayment term, and other relevant factors?

3. Lender behavior: What motivates lenders to fund loans on Kiva's platform, and what factors influence their decision to fund specific loans? Can we identify patterns or trends in lender behavior, such as preference for particular sectors, regions, or borrower demographics?

4. Field Partner efficiency: How do Field Partners' operational and financial characteristics affect loan distribution, borrower support, and repayment rates? Can we identify best practices among Field Partners to improve the overall effectiveness of Kiva's network?

5. Funding time analysis: What factors influence the time it takes for loans to be fully funded? Can we identify patterns in funding speed, and develop strategies to optimize loan visibility and promote faster funding for borrowers in need?

**Where is the data?** [Shared drive](https://drive.google.com/drive/folders/1ogLhA9Z7_yoDWhh1HXO-aKpL6AHTG64L?usp=sharing). Note that the data is quite big (5GB), which may require additional cleaning or analysis steps.

**Can multiple groups work on this?** Yes, but no more than TWO GROUPS with clearly distinguishable research questions and variables. Coordinate a time to meet with the researcher, then follow up afterward with a short email expressing your interest in working on this project and explaining why you should be selected.

**How to get more info about the data?** Contact Dr. Moran Koren ([korenmor@bgu.ac.il](mailto:korenmor@bgu.ac.il)).

## eQuality in Quality: Differences in scientific quality in Google Search

Does a kid growing up in Bangladesh get the same quality of information online as a kid growing up in Israel? Online information is often thought of as universally accessible. However, numerous studies have shown that language barriers and the quality of information differ substantially across the globe. In this project, we are interested in reexamining this question with respect to basic scientific terms and concepts (e.g. Atom), which should not differ as much from one language to another. We have collected the Search Results pages for 36 carefully selected scientific queries in multiple languages and locations around the world. We would like to understand the quality of content offered in each of these languages, how well search ranking works, and ultimately develop better measures for the quality of scientific search results. A stretch goal may include scraping web pages and training machine learning models to identify markers of scientific quality.

**Where is the data?** Contact Dr. Nir Grinberg ([nirgrn@bgu.ac.il](mailto:nirgrn@bgu.ac.il)).

**Can multiple groups work on this?** Yes, but no more than TWO GROUPS with clearly distinguishable research questions and variables. Coordinate a time to meet with the researcher, then follow up afterward with a short email expressing your interest in working on this project and explaining why you should be selected.

**How to get more info about the data?** Contact Dr. Nir Grinberg ([nirgrn@bgu.ac.il](mailto:nirgrn@bgu.ac.il)).

## AI-generated Email Reply Suggestions

Human-to-human communication is no longer just mediated by computers – it is increasingly generated by them, including on popular communication platforms such as Gmail, Facebook Messenger, Linkedin, and others. Yet, little is known about the differences between human- and machine-generated responses in complex social settings.

[We know](https://drive.google.com/file/d/1VPO6DWNPwvN8otVWzuMvJOBYffu6yj1Y/view?usp=drive_link) that Google Smart Reply suggestions (used by millions of people) could align more closely with human replies in terms of when responses should be offered, response length, sentiment, and semantic meaning. We [also know](https://dl.acm.org/doi/pdf/10.1145/3411764.3445557) that people find many of the reply suggestions inappropriate or unusable. Our work also demonstrated that newer generation models like GPT-3 could do better in some aspects, albeit providing looser guarantees for quality or safety.

There are two datasets available here:

1. EnronSR: an annotation of the Enron corpus with Smart Reply (SR) suggestions so we have for a set of ~34k emails both the human and SR responses (if there were any).
2. A survey of 240 people around the world who provided replies to the same set of 20 pre-selected emails, including people's demographics and emails in four communication contexts (work, family, friends, and acquintances).

Possible research questions:

1. Benchmarking: How do different LLMs compare in various dimensions (response rate, length, sentiment, semantics)? This would require *collecting* responses from multiple LLMs / APIs and analyzing their textual responses.
2. Bias in Smart Reply: Does SRs match more closely the responses of men or women? Rich or poor? Americans or Indians?
3. Analysis of variance in human responses to the same emails, perhaps along dimensions of communication goals.

**Where is the data?** Contact Dr. Nir Grinberg ([nirgrn@bgu.ac.il](mailto:nirgrn@bgu.ac.il)).

**Can multiple groups work on this?** Yes, but no more than TWO GROUPS with clearly distinguishable research questions and variables. Coordinate a time to meet with the researcher, then follow up afterward with a short email expressing your interest in working on this project and explaining why you should be selected.

**How to get more info about the data?** Contact Dr. Nir Grinberg ([nirgrn@bgu.ac.il](mailto:nirgrn@bgu.ac.il)).

## **Substance abuse: admission and discharge records**

Can we predict who will complete substance abuse treatment solely based on admission data? How does that vary across drugs and demographic groups? These are only some of the questions that could be explored using this public dataset.

TEDS-D is a national data system of annual discharges from substance use treatment facilities. A sibling data system—Treatment Episode Data Set: Admissions (TEDS-A)—collects data on admissions. TEDS-D contains records on admissions of people aged 12 and older, and includes information on admission demographics (for example, age, sex, race/ethnicity, employment status) and substance use characteristics (for example, substances used, age at first use, route of use, frequency of use, number of prior admissions).

**Where is the data?** <https://www.datafiles.samhsa.gov/dataset/treatment-episode-data-set-admissions-2020-teds-2020-ds0001>

**Can multiple groups work on this?** Yes, but no more than TWO GROUPS with clearly distinguishable research questions and variables.

**How to get more info about the data?** Read online about the dataset (and possibly other datasets available from SAMHSA).

# Additional data sources

Below are a list of data repositories that might be of interest to browse. You’re not limited to these resources, and in fact, you’re encouraged to venture beyond them. But you might find something interesting there:

* [Data.gov.il](https://data.gov.il/)
* [WRDS](https://wrds-www.wharton.upenn.edu/register/bgunegev/) (register with your @post.bgu.ac.il or @bgu.ac.il email only)
* [TidyTuesday](https://github.com/rfordatascience/tidytuesday)
* [NHS Scotland Open Data](https://www.opendata.nhs.scot/)
* [Edinburgh Open Data](https://edinburghopendata.info/)
* [Open access to Scotland’s official statistics](https://statistics.gov.scot/home)
* [Bikeshare data portal](https://www.bikeshare.com/data/)
* [UK Gov Data](https://data.gov.uk/)
* [Kaggle datasets](https://www.kaggle.com/datasets)
* [OpenIntro datasets](http://openintrostat.github.io/openintro/)
* [Awesome public datasets](https://github.com/awesomedata/awesome-public-datasets)
* [Youth Risk Behavior Surveillance System (YRBSS)](https://chronicdata.cdc.gov/Youth-Risk-Behaviors/DASH-Youth-Risk-Behavior-Surveillance-System-YRBSS/q6p7-56au)
* [PRISM Data Archive Project](https://www.icpsr.umich.edu/icpsrweb/content/ICPSR/fenway.html)
* [Harvard Dataverse](https://dataverse.harvard.edu/)
* https://snap.stanford.edu/data/index.html
* If you know of others, let me know, and we’ll add here…

If you are using a dataset that comes in a format that we haven't encountered, make sure that you are able to load it into R as this can be tricky.

\* Do not reuse datasets used in examples, homework assignments, or labs in the class.