1. Who are you (mini-bio) and what do you do professionally? If you are on a team, please complete this block for each member of the team.

Hossein Yousefi is a machine learning engineer with a focus on NLP and time series forecasting at Verto Health since January 2023. His journey in AI began in 2015 with a master's in computer vision for biomedical image analysis. Over the years, he has contributed to startups, big tech companies, and pursued graduate studies in the same field at the University of Toronto.

Issac Chan is a Machine Learning Engineer at Verto where he leverages advanced machine learning techniques to create impactful healthcare solutions. He holds a Master of Philosophy (M.Phil.) in Mechanical and Automation Engineering, with a research focus on unsupervised learning and representation learning. His passion for data science and commitment to staying abreast of cutting-edge research enable him to develop scalable, innovative solutions tailored to the healthcare sector.

Cho Yin Yong is an Engineering Manager at Verto Health and a sessional lecturer at University of Toronto teaching undergraduate Software Engineering courses. At Verto, he leads research and development on population health analytics. Recently, he led his team to work on a new NLP module that takes unstructured clinical notes for each patient, processes it into an International Patient Summary and aggregate it into population journeys.

2. What motivated you to compete in this challenge?

Patient stories are rarely documented as part of the patient chart. As one research study puts it: “The stories of people who attempt suicide are insufficiently reflected in suicide research in psychology” (Rimkeviciene 2016). In fact, according to many studies, including the NVDRS coding guide, many studies focused on only the most recent events that happened at most one month ago. Two excerpts of recent research papers are shown:

“About half of people who die by suicide visit their primary care provider (PCP) within 1 month of doing so, compared with fewer than 1 in 5 contacting specialty mental health.” (Dueweke 2018)
“...suicide-related outcomes within one-week or one-month in individuals with current suicidal ideation (SI) or a recent suicide attempt (SA).” (Lengvenyte 2021)

To better guide suicide prevention, we must first be informed of the series of events that victims gone through days, weeks or even months prior to death. We hope to inspire future research to look into the patient story with a broader timeframe as well as population timeline trends to more effectively prevent suicides.

3. High level summary of your approach: what did you do and why?

Step 0 - Remove notes without temporal information. Not all narratives were used in this research. In the current dataset of 4000 narratives, we kept **3201** narratives that contained potential temporal variables. We used simple string matching on common temporal representations, such as “month”, “day”, “ago”, etc.

Step 1 - Running valid notes through many rounds of Flan T5

- **Temporal Extraction**: Sequential Q&A through 3 flan-t5-xl prompts to segment the narrative into sentences and construct a structured temporal concept like `{'number': '1', 'unit': 'hour', 'before_or_after': 'before'}`
- **Sentence Topic Modeling**: We classified sentences into predetermined categories based on the existing boolean variables. For example, “V had just broken up with his girlfriend…” is classified as “relationship problem with partner” (`IntimatePartnerProblem`).

Step 2 - Event Log Creation

- **Filter narratives without temporal information**: 502 notes did not contain valid information or only contained the death with temporal variables. Those were removed from further processing. **2699** notes remained for final analysis.
- **Relative timing calculation heuristic function**: Each free-text temporal variable is now reformatted into an integer representing the relative hours prior to death. For example, “2 days ago” is reformatted into -48 (hours) and “2 months prior” is -1460 (hours)

The output of the process is an event log that can be used for data analysis purposes.

Final Result - Sankey Diagram from Event Log: We aggregated timelines per victim and constructed a Sankey diagram. Each node represents a significant event, such as relationship problem, suicide attempt. The transitions contain the median time to move to the next state, as well as the count of patients that moved to the next state. The last state is “Death of victim”.

For development, We used flan-t5-large for inference locally, and flan-t5-xl for the final run. We opted to use the `Standard_NC24ads_A100_v4` on Azure (80GB vRAM) for speed purposes when running the final pipeline. The main Python libraries used are: `nltk` for sentence segmentation, `transformers` for T5 inference, and `plotly, matplotlib` for visualization.

Why? Creating a population journey requires an event log as input. By **turning free-text into a time-series format**, we unlock the data-rich sources which are previously unavailable to the regular data analyst and data scientist to perform downstream tasks such as determining a patient’s most likely next step and present professionals with suitable interventions.

4. Please provide the machine specs and time you used to run your model.

Step 1:

Aggregate and extracted the time-related sentence data, follow with sentence classification for each victim

- Azure ML Studio Standard_NC24ads_A100_v4 (80GB vRAM)
- Inference duration: 2.5 hours

Step2:

Relative timing word extraction, normalization and classification.

- Azure ML Studio Standard_NC24ads_A100_v4 (80GB vRAM)
- Inference duration: 1.5 hours

All other steps

- Macbook Pro
- CPU: Apple M1 Pro
- RAM: 32GB
- Inference duration: within 2 hours
- Manually annotating data: 7 hours

5. Anything we should watch out for or be aware of in using your model (e.g. code quirks, memory requirements, numerical stability issues, etc.)?

At times, the LLMs that we use may hallucinate. Specifically, we found that predictions that showed more than 2 years delta was highly inaccurate, and it was excluded from analysis.

6. Did you use any tools, data, or pre-trained weights for data preparation or exploratory data analysis that aren’t listed in your code submission?

We manually annotated 40 samples for the ground truth for data preparation. Since the annotations contain the original narratives, it is not included in our code submission.

7. How did you evaluate the performance of your approach, if at all?

We used 7 hours to manually annotate 40 samples as the ground truth and compared them to the results generated by each step of our pipeline. Accuracy is calculated for all sentences if they are correctly extracted, but if there are false positives or missed sentences, it is calculated only for the correctly extracted ones.

Sentence extraction F1 score	Topic modeling accuracy	Temporal extraction accuracy
0.97	0.87	0.88

8. What are some other things you tried that didn't necessarily make it into the final workflow (quick overview)?

We explored using large language models such as the llama-3.2 and Qwen2.5 families. Smaller models in the above families had questionable output accuracy, while large models were cost-prohibitive and time-consuming. GLiNER (Zaratiana 2023) for NER was also attempted but we ultimately landed on the flan-t5-xl (Chung 2024) as it provided the most promising results.

Furthermore, we decided that a Sankey diagram output is more visually understandable than running it through heuristic mining, which was originally planned. In the end, we adhered to the "keep it simple" principle.

9. If you were to continue working on this problem for the next year, what methods or techniques might you try in order to build on your work so far? Are there other fields or features you felt would have been very helpful to have?

Ingesting more data: Only 2699 notes were possible to be analyzed in this research. Nonetheless, our pipeline supports more than just mental health narratives, as it was built to process any free-text clinical notes. Primary care and specialty care notes, emergency department discharge summaries, and even patient's personal diaries may allow for more detailed patient timelines.

More useful visualizations: Sankeys are one way to interpret time-series data in a human-readable format, but as our research outputs events in an event log format, the possibilities are endless. If more time was given, our next step would be to create a hidden Markov model based on existing data, popular in the clinical data science domain.

Better labelled ground truth: In this research, we only used pretrained models instead of finetuning our own, due to the lack of ground truth of temporal variables (that is understandable - it is a new variable we are proposing!) If we had more time, we could label more ground truth either manually or with the assistance of a larger LLM for further finetuning.