Generating and evaluating simulated medical notes: Getting a Natural Language Generation model to give you what you want

Robert Horton
rhorton@microsoft.com
Microsoft
Mountain View, CA, USA

Maryam Tavakoli Hosseinabadi
matavako@microsoft.com
Microsoft
Mountain View, CA, USA

Alexandre Vilcek
alvilcek@microsoft.com
Microsoft
Bellevue, WA, USA

Wolfgang M. Pauli
wopauli@microsoft.com
Microsoft
Bellevue, WA, USA

Mario E. Inchiosa
marinch@microsoft.com
Microsoft
Mountain View, CA, USA

ABSTRACT

Strong restrictions on sharing healthcare data pose a significant barrier to developing and applying machine learning (ML) technologies in this field. Significant effort has been invested in generating "realistic but not real" Electronic Medical Record (EMR) data that can be used to facilitate many aspects of the digital transformation effort in healthcare [17]. Here we demonstrate a transformer-based Natural Language Generation approach to supplement the structured EMR data produced by the open-source Synthea^{TM} simulation system with narrative text fields ('History of present illness') that are semantically consistent with the structured attributes for a given simulated patient encounter. One central hyperparameter for text generation is top_p, which determines the trade-off between diversity of generated text, while maintaining fluency and coherency. We evaluate the generated text via BERT-based text classification, regular expression matching, domain-specific entity recognition, and semantic embedding for repetition detection and study the impact of top_p on these metrics. Our observations show that increasing top_p improves some quality measures while worsening others. Input from domain experts will be required to find an optimal top_p for a specific task. This is preliminary work toward a larger goal of producing simulated text data suitable for developing and demonstrating various NLP-based ML approaches in EMR applications.

CCS CONCEPTS

• **Computing methodologies** → *Neural networks*; **Natural language generation**; • **Security and privacy** → **Human and societal aspects of security and privacy**.

KEYWORDS

natural language processing, natural language generation, responsible artificial intelligence, document intelligence

1 INTRODUCTION

Natural Language Generation (NLG) models such as OpenAI's GPT-3 [3], and Microsoft's Turing-NLG [13] have the ability to reasonably extrapolate additional text given a partial document "prompt," and can address an impressively wide range of Natural Language Processing (NLP) problems in this way. However, the computational expense involved in adapting and running these models often prevents us from using them directly; instead, we may need to distill behavior from a large NLG model into a smaller (and cheaper) model for production. One way to do this is to use the NLG model to generate (or augment) training data that in turn can be used to train conventional machine learning (ML) models, such as text classifiers or named entity recognition (NER) systems. This paper documents our efforts to generate and evaluate simulated *History of Present Illness* (HPI) narrative text fields to accompany simulated structured electronic medical record (EMR) data.

1.1 NLG as a Software Paradigm

In a new paradigm of software creation ("Software 3.0"[1]) you show the machine what you want it to do in the form of a prompt. The prompt is a partial document, and the NLG model's job is to continue where the prompt leaves off.

How can you tell if the model has generated something appropriate for your task? The challenge is to balance creativity with control; we want the machine to enrich our data with appropriate descriptive text (like mentioning that a broken arm is accompanied by pain and swelling or providing a backstory on how the patient broke her arm) while including specific structured facts consistent with the context of the simulated medical encounter. In an open-ended text generation setting, metrics such as ROUGE [9] or BLEU [12] are not useful [6], and in the absence of a ground truth, metrics such as perplexity [6] are not measurable. Therefore, we use four other approaches to analyze the quality of generated passages.

1.2 Simulated Medical Histories

Synthetic data is of particular interest in the medical domain [17]. The unique characteristics of medical data as well as privacy concerns about sharing patient information create barriers for collaboration between the clinical and data analysis communities. Though there is extensive literature on generating synthetic tabular [4, 10, 14, 17] and imaging [1, 7] data, simulating the free-text portions of medical records remains a challenge. Much of the previous work has used rule-based and machine learning methods to generate de-identified versions of actual clinical notes[2] [11]. Here

[1] Chris Olah Twitter thread https://twitter.com/ch402/status/1273765062633639936
[2] https://github.com/orenmel/synth-clinical-notes

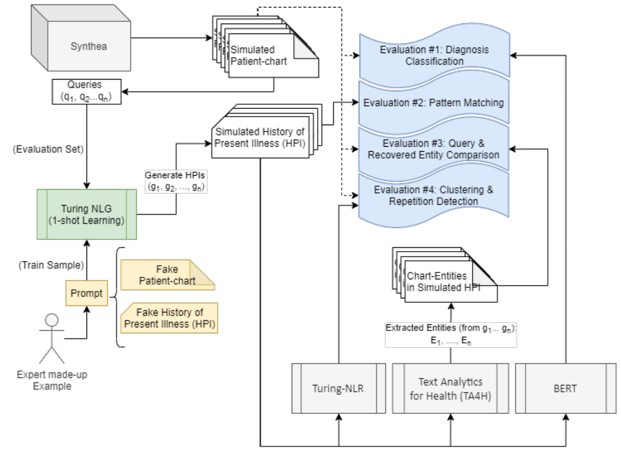

Synthea
Simulated Patient-chart
Queries (q_1, q_2...q_n)
(Evaluation Set)
Turing NLG (1-shot Learning)
Generate HPIs (g_1, g_2, ..., g_n)
Simulated History of Present Illness (HPI)
Evaluation #1: Diagnosis Classification
Evaluation #2: Pattern Matching
Evaluation #3: Query & Recovered Entity Comparison
Evaluation #4: Clustering & Repetition Detection
(Train Sample)
Prompt
Fake Patient-chart
Fake History of Present Illness (HPI)
Expert made-up Example
Chart-Entities in Simulated HPI
Extracted Entities (from g_1... g_n): E_1, ..., E_n
Turing-NLR
Text Analytics for Health (TA4H)
BERT

Figure 1: Experiment Design

we describe our preliminary work toward generating narrative text fields to accompany records generated by the Synthea^{TM} EMR simulation system. Synthea^{TM} is an agent-based platform supporting a large number of domain expert-contributed modules that make use of publicly available health statistics, clinical guidelines, and patient care protocols to model incidence, progression, and treatment of clinical conditions. The simulation runs over the course of each patient's lifetime, producing a population of patients in parallel. The resulting fully synthetic longitudinal medical records have shown to be to be suitable for a variety of nonclinical secondary uses, including education and many aspects of healthcare IT innovation [2, 16]. We use tabular data generated by Synthea to provide structured facts to the NLG system so it can incorporate them into narrative text.

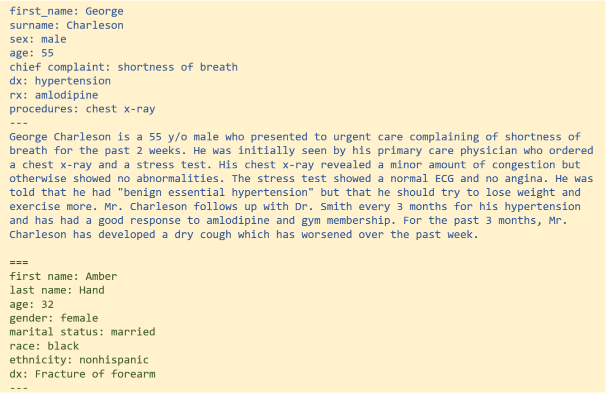
2 EXPERIMENT SETUP

The primary goal of this study is to generate HPI narratives for a set of simulated patient charts. The expectation is to get some narratives of high linguistic quality that are consistent with the patients' charts and the reason for that particular visit.

NLG involves sampling from probability distributions. This is typically done via nucleus sampling [8], which samples from the smallest vocabulary that covers the top_p portion of the probability mass. Choosing a value of p less than 1 causes generated text to be more coherent, by avoiding sampling from long tails of low-probability tokens. We will scan over Turing-NLG's `top_p` hyperparameter, to evaluate its effect on metrics of diversity and consistency. A summary of the data generation and evaluation flow is demonstrated in Figure 1.

2.1 Prompt design

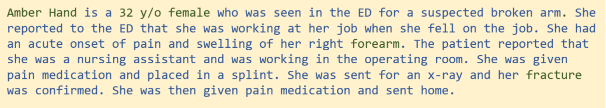
The prompt contains a single example HPI passage, preceded by a set of structured facts about the case; these facts are incorporated in the text. This example was used in the prompt for all patient encounters. It is appended with a new set of structured facts specific to each encounter. The NLG model completes the prompt by filling in text that (hopefully) incorporates these facts.



```
first_name: George
surname: Charleson
sex: male
age: 55
chief complaint: shortness of breath
dx: hypertension
rx: amlodipine
procedures: chest x-ray
---
George Charleson is a 55 y/o male who presented to urgent care complaining of shortness of
breath for the past 2 weeks. He was initially seen by his primary care physician who ordered
a chest x-ray and a stress test. His chest x-ray revealed a minor amount of congestion but
otherwise showed no abnormalities. The stress test showed a normal ECG and no angina. He was
told that he had "benign essential hypertension" but that he should try to lose weight and
exercise more. Mr. Charleson follows up with Dr. Smith every 3 months for his hypertension
and has had a good response to amlodipine and gym membership. For the past 3 months, Mr.
Charleson has developed a dry cough which has worsened over the past week.

===
first name: Amber
last name: Hand
age: 32
gender: female
marital status: married
race: black
ethnicity: nonhispanic
dx: Fracture of forearm
---
```

Figure 2: The *prompt* we used to induce Turing-NLG to generate a HPI narrative, having a few fields from the structured patient-chart information (highlighted in green).



```
Amber Hand is a 32 y/o female who was seen in the ED for a suspected broken arm. She
reported to the ED that she was working at her job when she fell on the job. She had
an acute onset of pain and swelling of her right forearm. The patient reported that
she was a nursing assistant and was working in the operating room. She was given
pain medication and placed in a splint. She was sent for an x-ray and her fracture
was confirmed. She was then given pain medication and sent home.
```

Figure 3: Example of generated HPI using the prompt in the previous figure. Entities directly matching structured facts from the prompt are highlighted in green.

2.2 Example outputs

Desired outputs should incorporate the information from the structured facts provided for that case and enrich it with general knowledge while not providing extraneous medical details, which could possibly contradict other details in the record. Figure 3 shows an example. In addition to the structured facts from the part of the prompt that was specialized for this particular patient encounter, the NLG model adds a variety of other information to fill out the story, including where the patient was seen, how she broke the arm, what general symptoms accompany this injury (pain and swelling), what tests are done (x-ray), etc. In future work, we will want to control more of these details, but we consider this example acceptable for the current iteration.

Some categories of undesired outputs, such as missing facts, can be recognized programmatically, while others will require expert judgment and will need to be addressed in future work. Here are some examples of undesired outputs: *"Mrs. Oretha Flatley is a 5 y/o female ..."* (This is a 5 year old girl who is married.) *"Melita is on the waiting list for an elbow replacement."* (Elbows are not like kidneys; you don't need to wait for someone to die to get one.) *"Ida Mares is a 15 year old female ..."* (Her name was supposed to be 'Isabela') *'... a 59 year old male, native Caucasian male ...'* (native Caucasian?) *'... was given a sling and crutches for her ankle.'* (You use a sling for a broken arm, not a broken ankle.)

Repeated phrases are a common error in transformer-based [15] NLG, especially for low values of top_p. Examples are shown in Figures 7 and 8; Section 3.4 addresses this kind of error.

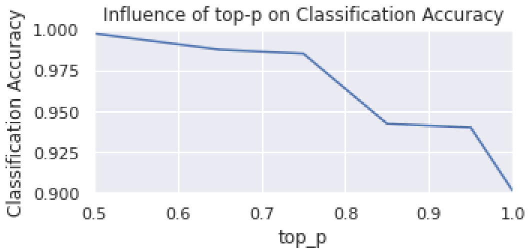


Figure 4: Classification accuracy versus `top_p` value. The high accuracy on the evaluation dataset suggests that most of the generated narratives are consistent with the corresponding diagnosis. Note that accuracy drops slightly as `top_p` increases, while other plots below show improvements with increasing `top_p` values.

3 EVALUATION OF OUTPUTS

3.1 Evaluation 1: Diagnosis classification

As a general evaluation of consistency between the generated narratives and their corresponding structured facts , we trained a BERT text classification model [3] [5] to predict the diagnosis from the generated text. We prepared a dataset with 6,000 generated narratives, each corresponding to one of two possible diagnoses: fracture of ankle and fracture of forearm. Figure 4 shows the influence of Turing-NLG's top_p parameter on the classification accuracy.

After analyzing the classification errors, we categorized the major types of inconsistencies found in the generated text. Here we list them and provide some examples:

(1) Facts described in generated text unrelated to any of the possible diagnoses (this is rare). Example text where the diagnosis in the corresponding structured facts is "fracture of forearm" and the model predicted "sprain of ankle":

> *"Jimmie Medhurst is a 67 y/o white male. He has a history of hypertension, hypercholesterolemia, gout, and atrial fibrillation. He is a retired former Army medic."*

(2) Facts in generated text related to one (or more) of the possible diagnoses, but don't completely describe it. An example where the diagnosis in the corresponding structured facts is fracture of ankle and the model predicted sprain of ankle:

> *"Mr. Lamar Runte is a 16 y/o male who was brought to urgent care with a splint on his left ankle. He claims that he injured his ankle when he slipped on the wet pavement. He states that his ankle has begun to feel tingly and is swollen to the point where it is difficult to walk. He also states that he was running late for work when he slipped. He was walking around at work when his boss noticed the injury."*

(3) Facts in the generated text describing more than one of the possible diagnoses. Example text where the diagnosis in the corresponding structured facts is fracture of ankle and the model predicted sprain of ankle:

> *"Marva Jacobson is a 16 y/o white female who has a history of ankle instability (hypoplastic bones in her feet). She presented to urgent care with a severe ankle sprain after falling in a parking lot. She complained of severe pain and swelling to her ankle. A radiograph of her ankle revealed that it had a moderate fracture. She is in a boot with crutches for the next 3 weeks."*

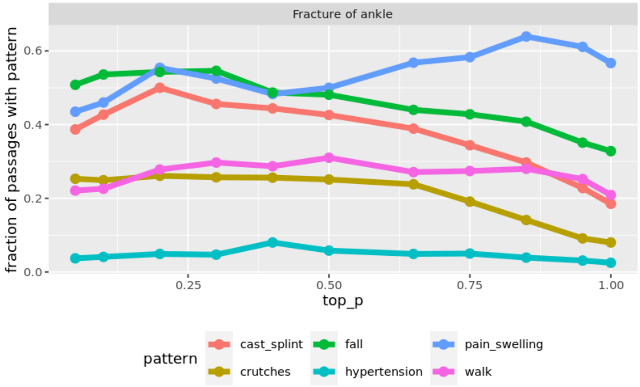


Figure 5: Fraction of passages containing specific regular expression patterns. Most of these terms are things we expect to find in a description of a patient with a broken ankle; these must be supplied by the model as general knowledge since they do not appear in the prompt. We also tracked the term "hypertension" since this concept does appear in the base prompt, though we did not intend for it to be incorporated in the results in general.

3.2 Evaluation 2: Pattern matching

We can use regular expression pattern matching to quantify specific results, such as the use of specific appropriate terminology, and plot the fraction of generated passages that match the pattern. Figure 5 shows the frequency of various matches at different values of top_p. We expect to see terms like "crutches" and "walk" in a discussion of a broken ankle, and words like "fall" in descriptions of how the ankle got that way. There is a low level of unintended leakage of details from the general prompt, exemplified by the term "hypertension" (this is not normally associated with broken arms, but it shows up in some of our passages presumably because it was in the prompt). Note that the frequency of carryover of "hypertension" seems to be affected by top_p. The same is true of "Dr. Smith"; when a physician's name is mentioned in our generated data, it is most commonly Dr. Smith, who was mentioned in the prompt.

[3] https://huggingface.co/bert-base-cased

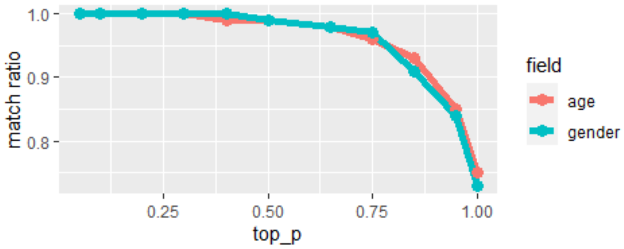
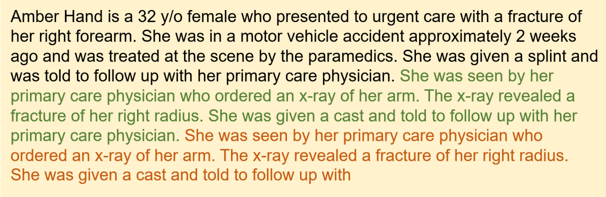


Figure 6: Exact-match ratio of the first gender/age entity of the generated text and the query. Though TA4H is not a perfect NER extractor, we assume that any retrieval errors are similar across all the samples.



Amber Hand is a 32 y/o female who presented to urgent care with a fracture of her right forearm. She was in a motor vehicle accident approximately 2 weeks ago and was treated at the scene by the paramedics. She was given a splint and was told to follow up with her primary care physician. She was seen by her primary care physician who ordered an x-ray of her arm. The x-ray revealed a fracture of her right radius. She was given a cast and told to follow up with her primary care physician. She was seen by her primary care physician who ordered an x-ray of her arm. The x-ray revealed a fracture of her right radius. She was given a cast and told to follow up with

Figure 7: Exact repeat detected with clustering at d=0.01. The passage ends mid-sentence because it reached the output token limit.

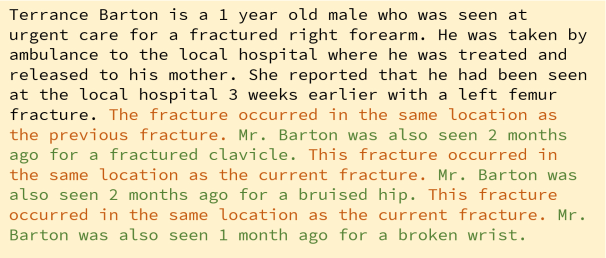
3.3 Evaluation 3: Comparing inputs to recovered entities

We used Microsoft's Text Analytics for Health (TA4H) [4], an Azure service for named entity recognition in the healthcare domain, and compared the retrieved entities for gender and age with the originally specified values. Figure 6 shows how the extracted entities match the demographic information in the prompt and the impact of top_p choice on the matching rate.

3.4 Evaluation 4: Clustering and repetition detection

We used the Turing AGIv5 encoder [18], called *Turing-NLR* in this paper, to generate semantic embeddings for sentences individually, then ran hierarchical clustering using cosine distance with the 'scipy.cluster.hierarchy' 'ward' method. Sentences were assigned to clusters using 'fcluster' with a distance threshold chosen by empirically examining clusters until we found one value (0.01) that gives groups of extremely similar (in most cases exactly repeated) sentences, and another value (0.1) that groups together sentences with clear similarity, while allowing some variability. This approach is useful for finding inexact repeats, as shown in Figures 7, 8, and 9.

The distinct passage fraction is the number of distinct passages divided by the total number of passages; we also compute the corresponding ratio at the sentence level. A distinct passage fraction of 1 means no passages were repeated within the run of 1000 passages, and a distinct sentence fraction of 1 means no sentences were repeated anywhere in the run of passages. A larger distance threshold requires that sentences (or passages) must be more different from one another to be considered distinct.



Terrance Barton is a 1 year old male who was seen at urgent care for a fractured right forearm. He was taken by ambulance to the local hospital where he was treated and released to his mother. She reported that he had been seen at the local hospital 3 weeks earlier with a left femur fracture. The fracture occurred in the same location as the previous fracture. Mr. Barton was also seen 2 months ago for a fractured clavicle. This fracture occurred in the same location as the current fracture. Mr. Barton was also seen 2 months ago for a bruised hip. This fracture occurred in the same location as the current fracture. Mr. Barton was also seen 1 month ago for a broken wrist.

Figure 8: Approximate repeats detected with clustering by loosening the cluster distance to d=0.1. Here two categories of near-repeats alternate.

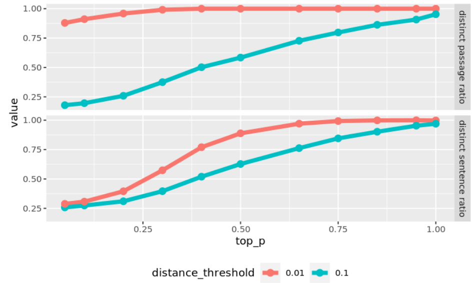


Figure 9: Fractions of sentences and passages that are distinct (not repeated), as a function of `top_p`.

4 DISCUSSION AND FUTURE DIRECTIONS

This paper investigated the impact of the hyperparameter top_p on measures of diversity and consistency in synthetic "history of present illness" notes, generated with the Turing NLG model. We assessed the quality of the generated notes using four different evaluation metrics and showed that classification accuracy and exact match ratio worsen with increasing top_p (Figures 4 and 7), while the fractions of sentences and passages that are distinct improve with increasing top_p (Figure 9). An optimal top_p value should reflect a balance between these opposing trends.

This work used simplified examples and did not follow a rigorous medical note structure. Our next iteration will require close collaboration with domain experts to design better prompts and develop structured rubrics for scoring how well the generated text meets criteria for correctness and realism, and how well it fits in with the simulated medical record. Medical experts will label examples for training and testing ML classifiers that can perform this kind of scoring. To optimize the use of the physicians' limited time, we plan to employ an active learning process for labeling. The present work on programmatic evaluation helps set the stage for this kind of domain expert supervision by weeding out some obvious problems, so our domain experts will be able to focus on the aspects that actually require their expertise.

[4] https://docs.microsoft.com/en-us/azure/cognitive-services/text-analytics/how-tos/text-analytics-for-health?tabs=ner

REFERENCES

[1] Saadia Binte Alam, Moazzem Hossain, and Syoji Kobashi. 2020. Synthetic Brain Image Generation for ADHD prediction based on Progressive Growing Generative Adversarial Network. In *International Symposium on Affective Science and Engineering ISASE2020*. Japan Society of Kansei Engineering, 1–5.

[2] Sandeep Bala, Angela Keniston, and Marisha Burden. 2020. Patient Perception of Plain-Language Medical Notes Generated Using Artificial Intelligence Software: Pilot Mixed-Methods Study. *JMIR Formative Research* 4, 6 (2020), e16670.

[3] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[4] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. 2017. Generating multi-label discrete patient records using generative adversarial networks. In *Machine learning for healthcare conference*. PMLR, 286–305.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[6] Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833* (2018).

[7] Amirata Ghorbani, Vivek Natarajan, David Coz, and Yuan Liu. 2020. DermGAN: synthetic generation of clinical skin images with pathology. In *Machine Learning for Health Workshop*. PMLR, 155–170.

[8] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751* (2019).

[9] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[10] Pei-Hsuan Lu, Pang-Chieh Wang, and Chia-Mu Yu. 2019. Empirical evaluation on synthetic data generation with generative adversarial network. In *Proceedings of the 9th International Conference on Web Intelligence, Mining and Semantics*. 1–6.

[11] Oren Melamud and Chaitanya Shivade. 2019. Towards Automatic Generation of Shareable Synthetic Clinical Notes Using Neural Language Models. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics, Minneapolis, Minnesota, USA, 35–45. https://doi.org/10.18653/v1/W19-1905

[12] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

[13] Corby Rosset. 2020. *Turing-NLG: A 17-billion-parameter language model by Microsoft*. Retrieved May 20, 2021 from https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/

[14] Matthias Templ, Bernhard Meindl, Alexander Kowarik, and Olivier Dupriez. 2017. Simulation of synthetic complex data: The R package simPop. *Journal of Statistical Software* 79, 10 (2017), 1–38.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762* (2017).

[16] J Walonoski, S Klaus, E Granger, D Hall, A Gregorowicz, G Neyarapally, A Watson, and J Eastman. 2020. Synthea™ Novel coronavirus (COVID-19) model and synthetic data set. *Intelligence-based medicine* 1 (2020), 100007.

[17] Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. 2018. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association* 25, 3 (2018), 230–238.

[18] Hongfei Zhang, Xia Song, Chenyan Xiong, Corby Rosset, Paul N. Bennett, Nick Craswell, and Saurabh Tiwary. 2019. Generic Intent Representation in Web Search. (2019), 65–74. https://doi.org/10.1145/3331184.3331198