Do numbers matter? Types and prevalence of numbers in clinical texts

Rahmad Mahendra, Damiano Spina, Lawrence Cavedon, and **Karin Verspoor**
RMIT University
Melbourne, Australia
rahmad.mahendra@student.rmit.edu.au
{damiano.spina, lawrence.cavedon, karin.verspoor}@rmit.edu.au

Abstract

In this short position paper, we highlight the importance of numbers in clinical text. We first present a taxonomy of number variants. We then perform corpus analysis to analyze characteristics of number use in several clinical corpora. Based on our findings of extensive use of numbers, and limited understanding of the impact of numbers on clinical NLP tasks, we identify the need for a public benchmark that will support investigation of numerical processing tasks for the clinical domain.

1 Introduction

Numbers comprise a considerable amount of textual content and contribute substantially to conveying meaning in a range of domains including financial and scientific contexts. Targeted strategies for representing numbers have been shown to improve general literacy of language models (Thawani et al., 2021a). Numbers pose challenges for Natural Language Processing (NLP) due to their varied representations in text, as digits, words, or numerical expressions. This requires NLP models to handle ambiguity and context effectively in interpreting numerical information (Thawani et al., 2021b).

Numerical reasoning is crucial in the generative large language model (LLM) era because it underpins data-driven decision-making in many fields, including clinical, where accurate numerical insights are essential. LLMs often struggle with arithmetic operations and unit conversions, impacting their reliability in quantitative tasks. In the clinical domain, accurate numerical reasoning is vital for analyzing trial data, interpreting results, and making precise treatment recommendations, such as determining appropriate drug dosages based on statistical analyses of patient outcomes. For example, generating a report that states "The mean number of antihypertensive medication classes increased from 1.6 (95% CI, 1.4-1.8) at baseline to 2.2 (95% CI 2.0-2.4) at 6 months"[1] requires precise numerical reasoning.

In this paper, we characterize the numerical information in clinical NLP corpora. Through corpus analysis, we find that numerical information is frequent, but only a small portion is annotated or utilized. Our analysis identifies a number of issues concerning numeracy that need further attention from the clinical NLP research community.

2 Numerical strings and types

There is an assumption that numbers may be trivially extracted from text, as they consist of digit sequences or a finite set of numerals. However, numerical information can appear in various lexical surface and semantic contexts (Hanauer et al., 2019; Miok et al., 2023). We identify a multitude of number variants, and further define semantic categories for numerical information.

Numerical values can be expressed in many forms: (i) **Digit**, including integer ('3', '100,000'), float ('0.5'), and negative ('−1'), (ii) **Number with unit** ('100 mg', '160/90mmHg'), (iii) **Fraction**, written using the division symbol ('5/32') or with a special symbol (½), (iv) **Number range** ('from 0 to 2 years', '1969-77'), (v) **Numeral**, can be alphabetic numbers ('twenty-five', 'two') or combinations of numbers and words ('1 million', '3k'), (vi) **Number with Quantifier** ('>', 'less than', 'about'), (vii) **Percentage** is written either as '%' or 'percent'), (viii) **Roman numeral** ('iii', 'V').

Table 1 presents a summary of the prevalent types of numerical data, as well as examples of where each can be found in clinical texts.

3 Corpus analysis

Clinical documents are rich with diverse numerical data, expressed in different manners and contexts.

[1]This example is from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4311883/

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 409–415
August 16, 2024.

Type	Description	Instantiations in clinical texts
Cardinal	Used for counting or quantifying items in a set; total number of elements (Mirza et al., 2017).	number of participants sample size
Ordinal	Indicate the relative position or rank of an element within a series.	tumor stage
Measurement	A numerical value, typically accompanied by a unit, representing an attribute of a measured entity (Göpfert et al., 2022; Harper et al., 2021).	vital signs: body temperature, blood pressure, heart rate laboratory values: white blood cell count, hormone level, cholesterol
Temporal	Dates ('17 June 2024', '05/08/10'), times ('9pm', "two years ago"), and duration ("in an hour") (Tourille et al., 2017).	duration of intervention gestational age date ranges
Frequency	The number of times something occurs within a given interval.	medication dosage frequency
Proportion	A scaled quantity based on relative size	hospital readmission rate % group experiencing an outcome
Ratio	a comparison between two quantities	
Math	Numbers, variables, and operators in a mathematical statement such as a formula or probability; arithmetic operations as well as functions (Lu et al., 2023).	estimate of effect with confidence interval; p value
Non-numerical	Numerical values lacking number properties, e.g. as part of categorical data (identifiers) or or a named entities (e.g., COVID-19)	medical classification: disease code, pharmaceutical code

Table 1: Types of numerical information, and instantiations of each type in clinical text

To illustrate this, we empirically analyze four clinical NLP corpora. Our selection of corpora covers (i) various clinical NLP *tasks* — information extraction, information retrieval, natural language inference, and question answering — and (ii) *document types* — paper abstracts, clinical notes, and patient description narratives.

For each corpus, we present descriptive analysis. We count the frequency of numbers, estimated by how many digits and numerals occur in the text based on regular expression matching. We find that numbers are highly prevalent in each corpus. To understand contexts of number use, we sample and evaluate a few instances from each corpus qualitatively.

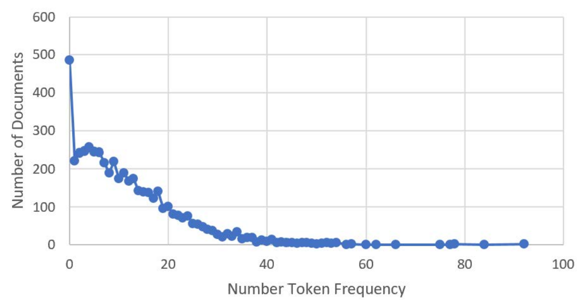

Figure 1: y EBM-NLP documents contain x numbers.

EBM-NLP (Nye et al., 2018) This corpus contains 4,993 abstracts of medical articles describing clinical randomized controlled trials in which text spans are annotated with PICO elements. That is, annotation labels include the trial (P)articipants enrolled, the (I)nterventions studied and to what they were (C)ompared, and the (O)utcomes measured. We find that 4,507 abstracts (90%) contain numerical information. The distribution of number token frequency with respect to number of abstracts is shown in Figure 1. The majority of abstract documents encode 5–20 numbers per abstract. However, only 13% of number tokens in the document collection are within annotated PICO-spans. About two-third of the numbers within spans belong to Participants entities, most of which relate to sample size and age of study population.

TREC-CDS (Koopman and Zuccon, 2016; Roberts et al., 2022) The TREC Clinical Trial series task involves matching a given patient to relevant clinical trials. The task is framed as retrieval of clinical trial documents using a patient descriptions as a topic query. The data includes 60 topics from

EBM-NLP (Nye et al., 2018)	
METHODS We obtained economic data from 1424 Guatemalan individuals (aged 25–42 years) between 2002 and 2004. They accounted for 60% of the 2392 children (aged 0-7 years) . . . enrolled in a nutrition intervention study during 1969–77. In this initial study, two villages were randomly assigned a nutritious supplement (atole) for all children and two villages a less nutritious one (fresco). . . . FINDINGS Exposure to atole before, but not after, age 3 years was associated with higher hourly wages, but only for men. For exposure to atole from 0 to 2 years, the increase was US$0.67 per hour (95% CI 0.16–1.17), which meant a 46% increase in average wages. There was a non-significant tendency for hours worked to be reduced and for annual incomes to be greater for those exposed to atole from 0 to 2 years.	
TREC-CDS (Roberts et al., 2022)	
Patient is a 55yo woman with h/o ESRD on HD and peritoneal dialysis who presented with watery, non bloody diarrhea and weakness. She has a history of 2 prior C diff infections, the most recent just 1 month ago. Recent antibx use in the last month on prior admission. Was also txd for Cdiff at that time for 14 d. course with po vanco. Pt was initially admitted to the ICU and was septic on pressors (levophed) until the morning of [**8–26**] with leukocytosis but no fever.	
MedNLI (Romanov and Shivade, 2018)	
Premise	The patient's hematocrit dropped from 29.7 to 22.8.
Hypothesis	The patient has a bleed.
Label	Entailment

Table 2: Sample instances involving numbers from EBM-NLP, TREC CDS, and MedNLI

TREC-CDS 2014 and 2015 (Koopman and Zuccon, 2016), 75 from TREC-CDS 2021, and 50 from 2022 (Roberts et al., 2022). We find that 100% of the patient descriptor topics contain numerical information. All of them contain patient age information expressed through different lexical variants. Most topics also contain numerical information about patient's vital signs, lab results, and medication history. Several types of numerical information are expressed as relations among measurement attributes, temporal information, and frequency.

MedNLI (Romanov and Shivade, 2018) Natural Language Inference (NLI) is an NLP task for determining whether a premise sentence semantically entails a hypothesis sentence. MedNLI is an NLI dataset sourced from clinical notes and annotated by a doctor. Each premise in MedNLI is grounded in the medical history of a patient and the hypothesis is a clinical conclusion labelled true, false, or maybe. We observe that nearly 50% of premise sentences contain numerical information, while only 1% of hypotheses have number tokens. This pattern is consistent across train, dev, and test data. We discover that numerical reasoning is one essential skill for formulating and interpreting medical conclusions.

PubMedQA (Jin et al., 2019) consists of a context and a yes/no/maybe question related to the context. Contexts are derived from PubMed abstracts, and questions are biomedical research questions. We find that 96.5% of contexts in the manually annotated subset PubMedQA-L contain numbers, including statistical information relating to trial results. Quantitative reasoning is needed to correctly infer the answer from the context.

4 Numeracy task and data in clinical domain

Thawani et al. (2021b) and Yoshida and Kita (2021) reviewed a broad range of numeracy tasks. Neither survey specifically considers the clinical domain. Given the findings of our corpus analysis (§3) that numbers are ubiquitous, we speculated that there may have been a number of related works on mining numerical information from clinical corpora. We search papers from the ACL Anthology[2] and PubMed[3] using the following keyword query.

> ("number" OR "numerical" OR "numeracy") AND ("clinical" OR "medical")

Among the numeracy tasks explored in the retrieved literature, mostly pertaining to information extraction, are extraction of lab test results (Bhatia et al., 2010; Liu et al., 2017) and extraction of measurement values from radiology report narrative (Bozkurt et al., 2019). In addition, we are aware that a number of clinical information extraction tasks also extract numerical attributes together with other entities, for example clinical trial variables (number of participants, sample size, outcome measurements) (Kiritchenko et al., 2010; Summerscales et al., 2011), eligibility criteria from clinical trials (Kury et al., 2020; Tseo et al., 2020),

[2] https://aclanthology.org/

[3] https://pubmed.ncbi.nlm.nih.gov/

medication attributes, such as drug dosage and frequency (Uzuner et al., 2010; MacKinlay and Verspoor, 2013; Kartchner et al., 2023), and temporal information (Sun et al., 2013; Styler IV et al., 2014; Miller et al., 2015)

Only a small number of tasks attempt to address numerical reasoning problems. One is NLI4CT, a dataset introduced for natural language inference and evidence retrieval tasks on clinical trial report (Jullien et al., 2023). Another addresses inference of patient phenotype based on extracted numerical values utilising one or more clinical attributes, e.g., "temperature 102°F suggesting Fever" (Tanwar et al., 2022).

5 Discussion

There are several possible directions to progress treatment of numbers in clinical NLP research.

Need for Benchmarks While there has been some research on extraction of numerical information from different medical data, most work has used their own data and the gold standard has not always been made public. This limitation has made it impossible to compare model performances of different systems (Jonnalagadda et al., 2015). The performance reported in several past works was very high (accuracy > 90%). This raises the question of whether numerical information extraction is a solved task.

We raise two concerns. First, a number of works utilized relatively small data and this may result in the reported accuracy scores lacking statistical significance. Second, some works applied 'easy' task formulations, i.e., given a sentence containing only a single number mention, it is trivial to extract most numerical attributes. Such spurious patterns in evaluation data may not generalize when we deal with more realistic scenarios for information extraction (Elangovan et al., 2024). For example, a sentence containing multiple numbers and more than one candidate for entities and attributes (see example of EBM-NLP instance in Table 2). Hence, we advocate for more public data benchmarks to transparently evaluate the progress of numerical information tasks in the clinical domain.

Scope of Numerical Reasoning Recent works on numerical reasoning deal with math and arithmetic problems (Mishra et al., 2022; Hendrycks et al., 2021; Cobbe et al., 2021). In fact, complex mathematics are less applicable in the clinical domain. In addition to arithmetic, other types of reasoning are required for clinical decision support. For example, number comparison (Park et al., 2022) and number normalization (Almasian et al., 2023). Different units of measurement often need conversion, requiring precise calculations to maintain accuracy. For example, hormone levels may be measured using nmol/L, ng/dL and ng/mL in different trials. Dealing with various number representation is important to interpret numerical information correctly. On the other hand, contextualizing such numbers with medical background knowledge is another important numeracy skill, as showcased by the example of MedNLI instance in Table 2.

Tokenization Challenges How to encode numbers in language models has been discussed in several works (Spithourakis and Riedel, 2018; Wallace et al., 2019; Thawani et al., 2021a). Encoding numbers is related to the problem of tokenization (Geva et al., 2020). The models struggle to recognize extrapolated numbers that are seldom found in corpora and force them to be tokenized digit by digit (Kim et al., 2021). In the clinical context, when the numbers are grounded with units or appear as ranges, the tokenizer is expected to be more robust. We inspected few samples of token-level annotated data of the EBM-NLP corpus. Our finding was that numbers are not fully correctly tokenized (e.g., "95% CI 0.16-1.17" is segmented into multiple individual digit numbers that lack meaning), even in the gold standard.

Utilizing Numerical Information for Clinical Application. Clinical documents contain numerous numerical data points. However, numbers are mostly neglected when designing an NLP system (Thawani et al., 2021b). Entity annotations skip numbers in most cases, as in the EBM-NLP corpus PICO annotation (Nye et al., 2018). Several clinical NLP works acknowledge the importance of numerical reasoning, but leave it for future work. For instance, in multi-document summarization, Otmakhova et al. (2022) identify that automatically generating systematic reviews involves meta-analysis that requires numerical aggregation of data across primary studies or calculating some statistics for variables. In another example, Lehman et al. (2019) argue that numerical information from the result section of studies can be utilized to improve evidence inference.

6 Conclusions

We analyzed well-established clinical NLP corpora, covering a variety of tasks and data sources, and identifying a broad set of types and usage of numbers. Our analysis shows that numbers play a major role in medical texts. On the basis of these findings, and the lack of systematic resources in the clinical domain for investigating numerical information extraction and reasoning tasks, we argue for the need for the construction of such resources. Numbers contain vital medical information. We strongly encourage clinical NLP researchers to consider how numerical processing may interact with their work.

Limitation

Our conclusion is based on the corpus we analyzed and reviewed during literature search. We may not include some corpora, especially those that are not publicly available, in our analysis. On the other hand, this work focuses only on English. While there have been some relevant works in the clinical domain for languages other than English, we leave this for future work.

Acknowledgments

The authors would like to acknowledge Country. This research has been carried out on the unceded lands of the Woi Wurrung and Boon Wurrung language groups of the eastern Kulin nation. We pay our respects to their Ancestors and Elders, past, present, and emerging. We respectfully acknowledge First Nations' connection to land, waters, sky, and community. This research was funded by the Australian Research Council through an Industrial Transformation Training Center Grant (grant IC170100030).

References

Satya Almasian, Vivian Kazakova, Philipp Göldner, and Michael Gertz. 2023. CQE: A comprehensive quantity extractor. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12845–12859, Singapore. Association for Computational Linguistics.

Ramanjot Singh Bhatia, Amber Graystone, Ross A Davies, Susan McClinton, Jason Morin, and Richard F Davies. 2010. Extracting information for generating a diabetes report card from free text in physicians notes. In *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*, pages 8–14, Los Angeles, California, USA. Association for Computational Linguistics.

Selen Bozkurt, Emel Alkim, Imon Banerjee, and Daniel L. Rubin. 2019. Automated detection of measurements and their descriptors in radiology reports using a hybrid natural language processing algorithm. *Journal of Digital Imaging*, 32(4):544–553.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Aparna Elangovan, Jiayuan He, Yuan Li, and Karin Verspoor. 2024. Principles from clinical research for NLP model generalization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, page to appear, Mexico City, Mexico. Association for Computational Linguistics.

Mor Geva, Ankit Gupta, and Jonathan Berant. 2020. Injecting numerical reasoning skills into language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 946–958, Online. Association for Computational Linguistics.

Jan Göpfert, Patrick Kuckertz, Jann Weinand, Leander Kotzur, and Detlef Stolten. 2022. Measurement extraction with natural language processing: A review. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2191–2215, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David A. Hanauer, Qiaozhu Mei, V. G. Vinod Vydiswaran, Karandeep Singh, Zach Landis-Lewis, and Chunhua Weng. 2019. Complexities, variations, and errors of numbering within clinical notes: the potential impact on information extraction and cohort-identification. *BMC Medical Informatics and Decision Making*, 19(3):75.

Corey Harper, Jessica Cox, Curt Kohler, Antony Scerri, Ron Daniel Jr., and Paul Groth. 2021. SemEval-2021 task 8: MeasEval – extracting counts and measurements and their related contexts. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 306–316, Online. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *NeurIPS*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the*

9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Siddhartha R. Jonnalagadda, Pawan Goyal, and Mark D. Huffman. 2015. Automating data extraction in systematic reviews: a systematic review. *Systematic Reviews*, 4(1):78.

Mael Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Dónal Landers, and Andre Freitas. 2023. NLI4CT: Multi-evidence natural language inference for clinical trial reports. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16745–16764, Singapore. Association for Computational Linguistics.

David Kartchner, Selvi Ramalingam, Irfan Al-Hussaini, Olivia Kronick, and Cassie Mitchell. 2023. Zero-shot information extraction for clinical meta-analysis using large language models. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 396–405, Toronto, Canada. Association for Computational Linguistics.

Jeonghwan Kim, Giwon Hong, Kyung-min Kim, Junmo Kang, and Sung-Hyon Myaeng. 2021. Have you seen that number? investigating extrapolation in question answering models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7031–7037, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Svetlana Kiritchenko, Berry de Bruijn, Simona Carini, Joel Martin, and Ida Sim. 2010. ExaCT: automatic extraction of clinical trial characteristics from journal publications. *BMC Med. Inform. Decis. Mak.*, 10(1):56.

Bevan Koopman and Guido Zuccon. 2016. A test collection for matching patients to clinical trials. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, page 669–672, New York, NY, USA. Association for Computing Machinery.

Fabrício Kury, Alex Butler, Chi Yuan, Li-heng Fu, Yingcheng Sun, Hao Liu, Ida Sim, Simona Carini, and Chunhua Weng. 2020. Chia, a large annotated corpus of clinical trial eligibility criteria. *Scientific Data*, 7(1):281.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. 2019. Inferring which medical treatments work from reports of clinical trials. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3705–3717, Minneapolis, Minnesota. Association for Computational Linguistics.

Sijia Liu, Liwei Wang, Donna M. Ihrke, Vipin Chaudhary, Cui Tao, Chunhua Weng, and Hongfang Liu. 2017. Correlating lab test results in clinical notes with structured lab data: A case study in hba1c and glucose. *AMIA Summits on Translational Science Proceedings*, pages 221 – 228.

Pan Lu, Liang Qiu, Wenhao Yu, Sean Welleck, and Kai-Wei Chang. 2023. A survey of deep learning for mathematical reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14605–14631, Toronto, Canada. Association for Computational Linguistics.

Andrew MacKinlay and Karin Verspoor. 2013. Information extraction from medication prescriptions within drug administration data. In *The 4th International Workshop on Health Document Text Mining and Information Analysis with the focus of Cross-language Evaluation (LOUHI), Canberra/Sydney, Australia*.

Timothy Miller, Steven Bethard, Dmitriy Dligach, Chen Lin, and Guergana Savova. 2015. Extracting time expressions from clinical text. In *Proceedings of BioNLP 15*, pages 81–91, Beijing, China. Association for Computational Linguistics.

Kristian Miok, Padraig Corcoran, and Irena Spasić. 2023. The value of numbers in clinical text classification. *Machine Learning and Knowledge Extraction*, 5(3):746–762.

Paramita Mirza, Simon Razniewski, Fariz Darari, and Gerhard Weikum. 2017. Cardinal virtues: Extracting relation cardinalities from text. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 347–351, Vancouver, Canada. Association for Computational Linguistics.

Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, and Ashwin Kalyan. 2022. LILA: A unified benchmark for mathematical reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5807–5832, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207, Melbourne, Australia. Association for Computational Linguistics.

Yulia Otmakhova, Karin Verspoor, Timothy Baldwin, Antonio Jimeno Yepes, and Jey Han Lau. 2022. M3: Multi-level dataset for multi-document summarisation of medical studies. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3887–3901, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sungjin Park, Seungwoo Ryu, and Edward Choi. 2022. Do language models understand measurements? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1782–1792, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kirk Roberts, Dina Demner-Fushman, Ellen M. Voorhees, Steven Bedrick, and William R. Hersh. 2022. Overview of the TREC 2022 clinical trials track. In *Proceedings of the Thirty-First Text REtrieval Conference, TREC 2022, online, November 15-19, 2022*, volume 500-338 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.

Georgios Spithourakis and Sebastian Riedel. 2018. Numeracy for language models: Evaluating and improving their ability to predict numbers. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2104–2115, Melbourne, Australia. Association for Computational Linguistics.

William F. Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, and James Pustejovsky. 2014. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

Rodney L. Summerscales, Shlomo Argamon, Shangda Bai, Jordan Hupert, and Alan Schwartz. 2011. Automatic summarization of results from clinical trials. In *2011 IEEE International Conference on Bioinformatics and Biomedicine*, pages 372–377.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *J. Am. Med. Inform. Assoc.*, 20(5):806–813.

Ashwani Tanwar, Jingqing Zhang, Julia Ive, Vibhor Gupta, and Yike Guo. 2022. Phenotyping in clinical text with unsupervised numerical reasoning for patient stratification. *Exp. Biol. Med. (Maywood)*, 247(22):2038–2052.

Avijit Thawani, Jay Pujara, and Filip Ilievski. 2021a. Numeracy enhances the literacy of language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6960–6967, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Avijit Thawani, Jay Pujara, Filip Ilievski, and Pedro Szekely. 2021b. Representing numbers in NLP: a survey and a vision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–656, Online. Association for Computational Linguistics.

Julien Tourille, Olivier Ferret, Xavier Tannier, and Aurélie Névéol. 2017. Temporal information extraction from clinical text. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 739–745, Valencia, Spain. Association for Computational Linguistics.

Yitong Tseo, M. I. Salkola, Ahmed Mohamed, Anuj Kumar, and Freddy Abnousi. 2020. Information extraction of clinical trial eligibility criteria. *CoRR*, abs/2006.07296.

Ozlem Uzuner, Imre Solti, and Eithon Cadag. 2010. Extracting medication information from clinical text. *J. Am. Med. Inform. Assoc.*, 17(5):514–518.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do NLP models know numbers? probing numeracy in embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5307–5315, Hong Kong, China. Association for Computational Linguistics.

Minoru Yoshida and Kenji Kita. 2021. Mining numbers in text: A survey. *Information Systems-Intelligent Information Processing Systems, Natural Language Processing, Affective Computing and Artificial Intelligence, and an Attempt to Build a Conversational Nursing Robot*.