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

**Rahmad Mahendra, Damiano Spina, Lawrence Cavedon, and Karin Verspoor**

**School of Computing Technologies - RMIT University, Melbourne, Australia**

ARC TRAINING CENTRE IN COGNITIVE COMPUTING FOR MEDICAL TECHNOLOGIES

RMIT UNIVERSITY

Link to Paper

**RCT Paper Abstract** PMID: 18242415

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). We used linear regression models ... to assess the relation between economic variables and exposure to atole ... at specific ages between birth and 7 years.

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.

**Patient Topic Descriptor**

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.

**Cardinal** **Money** **Time** **Age** **Percentage** **Math & Statistics** **Non-numerical**

## Characterizing Numbers

**Corpus Analysis**

**Semantic Types:**
- Cardinal
- Ordinal
- Measurement
- Temporal
- Money
- Frequency
- Proportion
- Ratio
- Math
- Non-numerical

**Number Taxonomy Construction**

**Lexical Form Variants**
- Digit
- Number with unit
- Fraction
- Number range
- Numeral
- Number with Quantifier
- Percentage
- Roman numeral

**Literature Review**
(ACL Anthology, PubMed)

## Findings

**EBM-NLP**
(Nye et.al., 2018)
- 4,057/4,993 abstracts (90%) contain numerical information
- 5-20 numbers per abstract
- < 15% of number tokens are within annotated PICO-spans

**TREC-CDS**
(Koopman and Zuccon, 2016; Roberts et al., 2022)
- 100% of topics contain numerical information (e.g., patient age), written in different lexical variants

**MedNLI**
(Romanov and Shivade, 2018)
- ~50% of premise sentences contain numerical information vs. 1% of hypotheses have number tokens
- Numerical reasoning is important

**MedNLI**

| | |
|---|---|
| **Premise** | The patient's hematocrit dropped from 29.7 to 22.8. |
| **Hypothesis** | The patient has a bleed. |
| **Label** | **Entailment** |

## Discussions

**Need for Benchmark**
- most existing gold standard is not publicly available
- small data -> statistical insignificant evaluation result
- 'easy' task formulation

**Scope of Numerical Reasoning**
- numerical reasoning math

**Number Representation and Tokenization challenge**

**Utilizing Numerical Information for Clinical Application**
- e.g., systematic review generation, evidence inference

## TLDR

**"Despite their importance in text, numbers are often overlooked in clinical NLP research."**