IT UNIVERSITY OF COPENHAGEN

[Citation needed] Data usage and citation practices in medical imaging conferences

Théo Sourget, Ahmet Akkoç, Stinna Winther,
Christine Lyngbye Galsgaard, Amelia Jiménez-Sánchez,
Dovile Juodelyte, Caroline Petitjean, Veronika Cheplygina

tsou@itu.dk
tsourget.fr

Introduction

- Papers mostly focus on method and not on data used
- Over-representation of some medical problems
- No citation standards for dataset usage, making detection difficult
- **How can we automatically quantify dataset usage?**

Tools

- Information from OpenAlex API, an open citation index tool
- Scraping of full text from MICCAI and MIDL
- Extraction of full-text data from PDFs with GROBID
- **Objective: detect the citations and mentions of 20 public medical datasets**

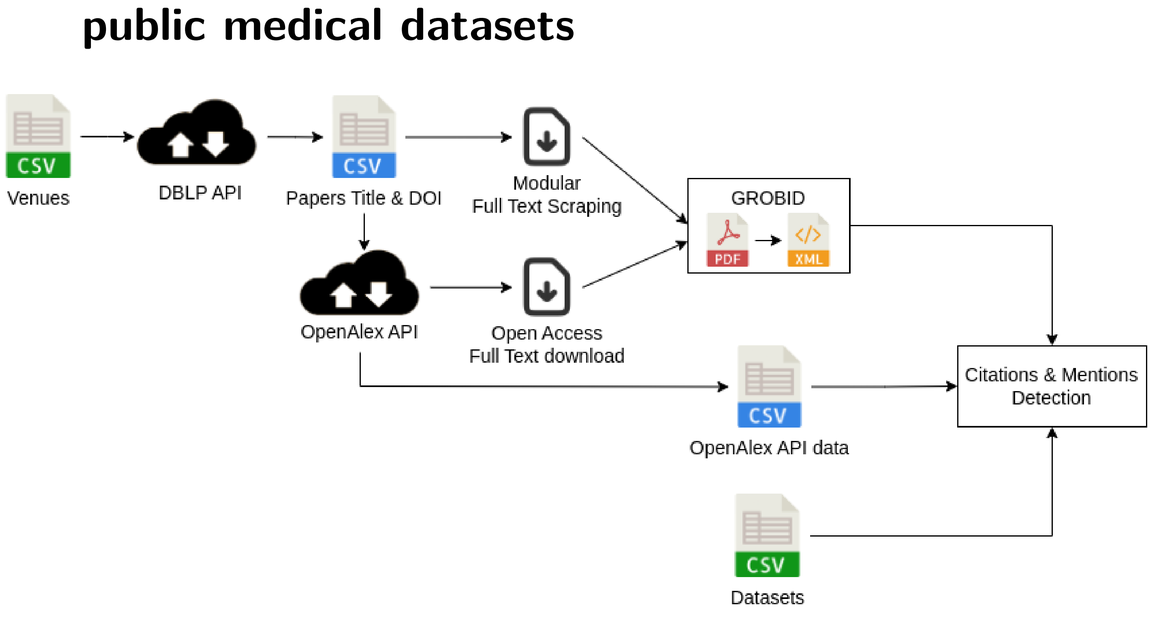

Figure 1: Flowchart of our method to quantify dataset usage in scientific papers. Blue CSV represents extracted data while green CSV represents manually selected data

Quantifying usage

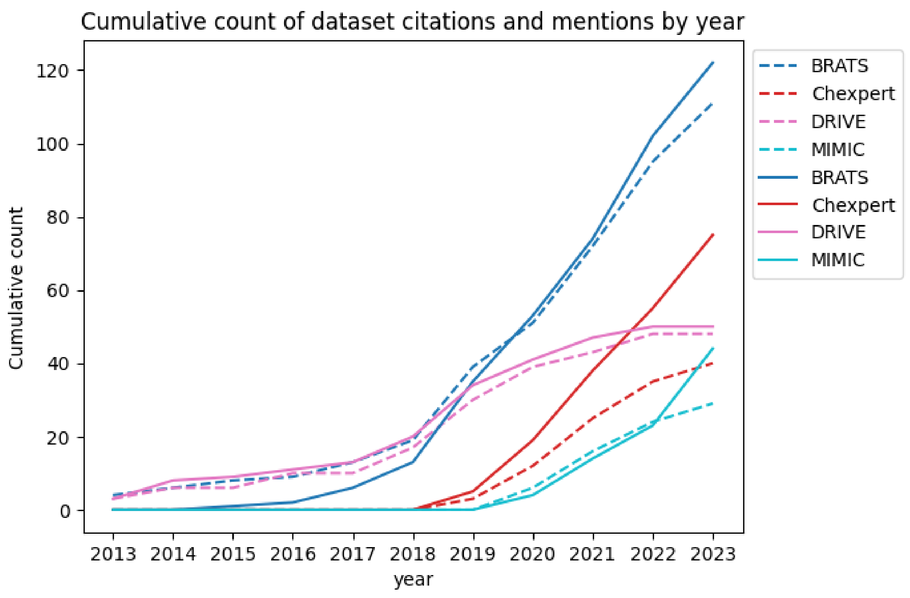

Figure 2: Cumulative counts per year of dataset citations (full lines) and mentions (dashed lines)

- Most datasets were almost never detected in papers
- Lack of diversity for the same task
- Old datasets are still highly cited
- Some factors like organizing a challenge or updating the dataset increase the snowball effect

Different citation practices

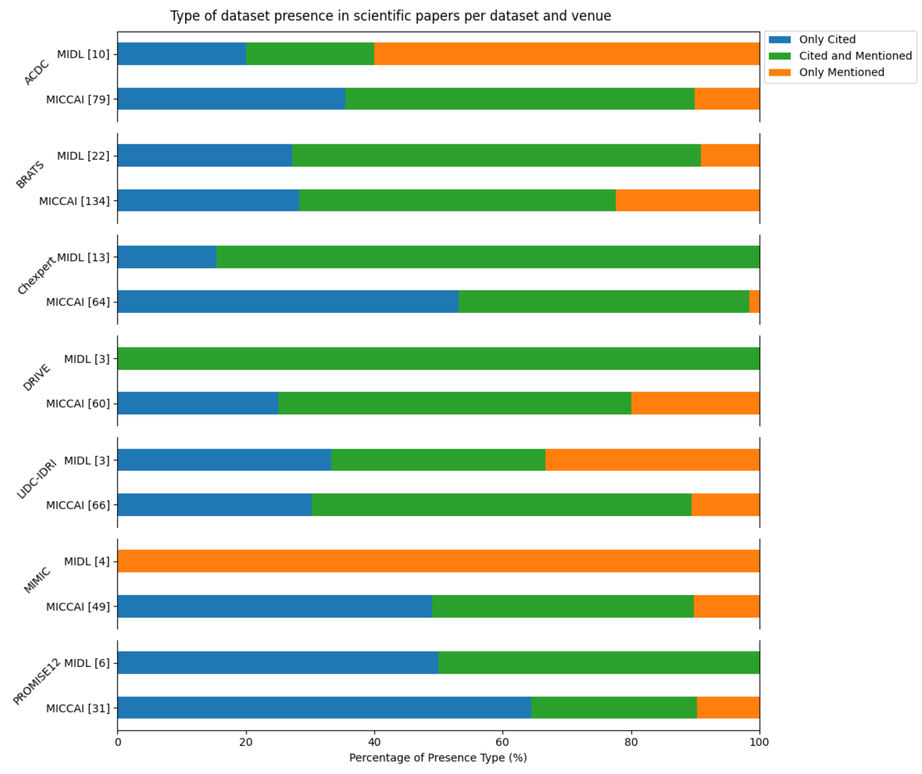

Figure 3: Type of presence per dataset and venue. The number in [] indicates the total number of papers for this subset

- "Only Cited": papers that are not using a dataset
- "Only Mentioned": papers with bad citation practice
- "Cited and Mentioned": papers with the ideal way to show usage

Discussion & Recommendations

- A real need for citation standard for dataset to track usage
- Improve awareness on dataset diversity in research papers
- An open and adaptable tool to quantify dataset usage across venues
- Another open-access tool for multiple users to label PDF documents

Acknowledgement

Danish Data Science Academy (DDSA) visit grant DDSA-V-2022-004 and DFF - Inge Lehmann 1134-00017B