SA83

Frequency and Type of Errors in Data Extraction within Systematic Literature Reviews

King E.[1], Roussi K.[1], Rice H.[1], Martin A.[1]

[1] Crystallise Ltd. Stanford-le-Hope, UK

Methodology

We analysed checked data extraction (DE) sheets from eight SLRs varying in topic and size conducted by our organization between 2022 and 2023.

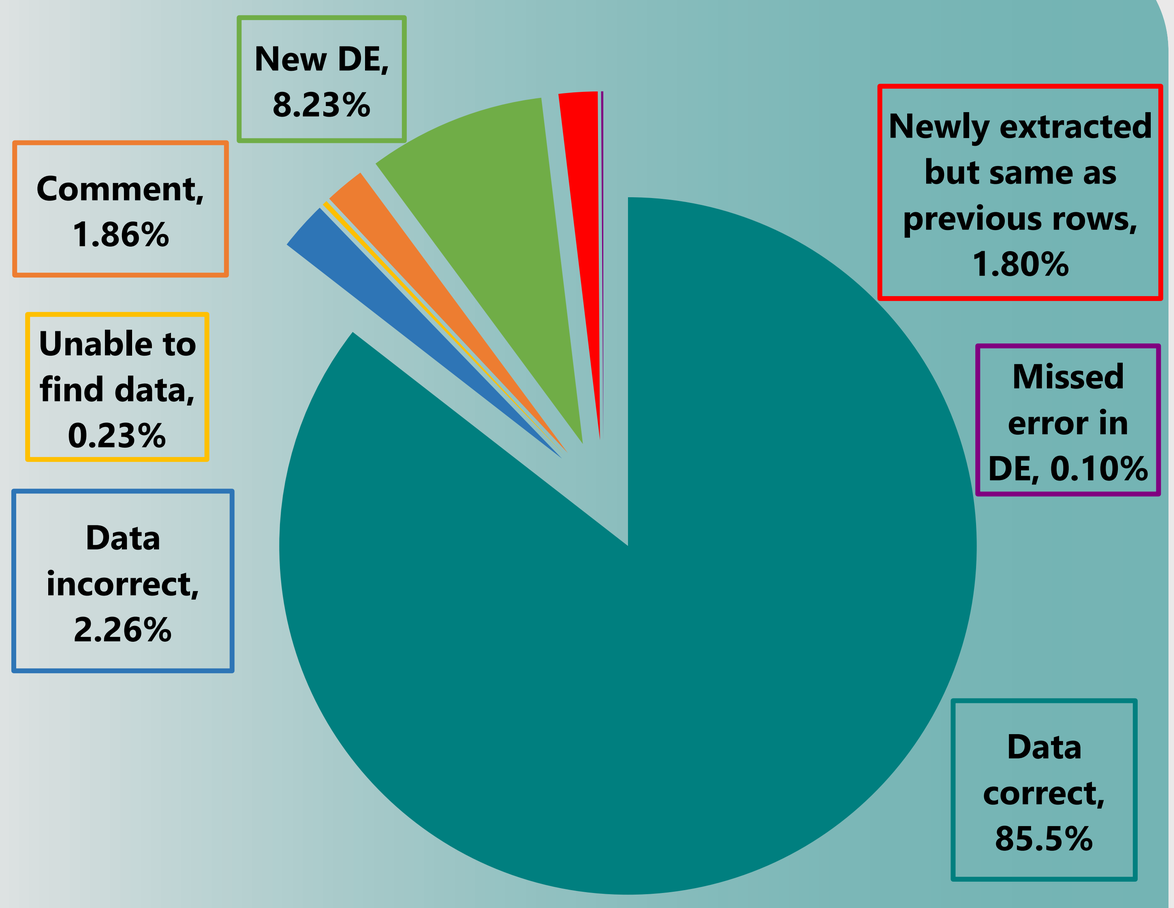
We calculated:

The **proportion of papers with errors** in each SLR

The **total number of errors** per paper and per project

The **different types of errors** per paper and per project

A score-based approach was devised to assess the difficulty of extraction, based on the publication type (full text/ abstract), whether the file was editable, whether it was highlighted ahead of DE, the number of pages and whether it was a new or updated SLR.



When the data was incorrect

60.3% Mismatch
Data were correctly extracted but were allocated to the wrong section of DE

23.2% Omission/Incomplete or Misidentification Relevant data not extracted

8.8% Misinterpretation Extracted or calculated data were incorrect for the allotted outcome

5.1% Typographical
Extracted data were incorrect due to typo or indistinguishable calculation error

2.7% Ambiguous
Extracted data was marked as incorrect without clear reason

Results

Initial data extraction by 13 different researchers was correct in **85.5% of 96,675 data points** evaluated.

In total, **59%** of papers included in all SLRs **had at least one error** at initial DE that was corrected during checking. The **most common error was misidentification (8.23%)**, when additional relevant data from the paper were identified by the checker, shown in grey in the figure above. Incorrect data, i.e., where the original value was incorrect, occurred in 2.26% of data points. Other changes were made to the DE by the checker in 3.89% of data points (e.g., inserting comments). Data misidentification (e.g., the correct value was inserted into the wrong column) occurred in 0.49% of data points.

No obvious pattern was found between the duration of DE or with the paper DE difficulty score and the DE error rate.

Discussion

Data extraction is an essential part of SLRs; however, it is error-prone. Other studies have identified DE error rates of 0.5% to 15% and at least one error in 66.8% to 99.3% of papers in published SLRs so the **>85% accuracy in overall data points** in our process **before pre-publication checking** compares favourably.

In future, to reduce data omissions, methods to **clarify all outcomes to be extracted** before DE starts should be explored as well as further in-depth analysis of the subtypes of errors in DE, such as the nature of mismatching or misinterpreted data.

Email: contact@crystallise.com

Website: www.crystallise.com

Crystallise

LinkedIn (Crystallise Ltd)

YouTube (@crystallise3499)