Superior Data Extraction for Systematic Literature Reviews using giles®

Ivan V.[1], Ionita T.[1], Veresteanu A.[1], Dasor V.[1], Shah N.[1], Liyanage G.[1], Raja S.[1], Wadhera R.[1], Khan K.[2], Aurangzeb M.[2], Khan I.[3]

[1]giles® AI, London, United Kingdom, [2]Regulatory Scientific and Health Solutions, [3]University of Warwick

giles® in collaboration with Google R-S-S Regulatory Scientific and Health Solutions

0.95 Overall Accuracy | **0.94** Reported Data Accuracy | **0.95** NR Precision | **0.98** NR Recall

Introduction

R-S-S (Regulatory Scientific and Health Solutions) is a functional service provider to the biopharmaceutical industry, supported by regulatory and technical experts who combine their knowledge of drug licensing with Health Technology Assessment (HTA).

Part of HTA (and regulatory) requirements includes evidence synthesis which requires systematic literature reviews (SLRs). R-S-S routinely generates SLRs across multiple therapeutic areas which are critical for the HTA process, as well as finding evidence gaps, designing future studies, and supporting submissions.

The manual data extraction performed during SLRs is a very time and resource intensive process. R-S-S have collaborated with giles® in bringing together a new AI-based tool that can facilitate accelerated SLRs, especially the data extraction process.

Objective

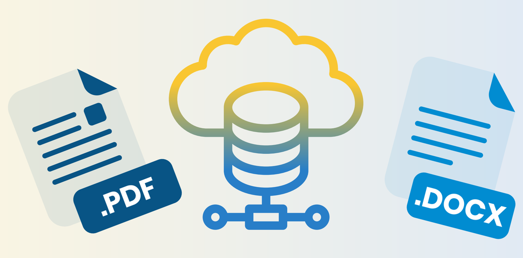
To build and evaluate a system that automatically extracts data from publications (typically found in PubMed amongst other online databases, including NICE HTA Assessment reports), given a set of pre-defined extraction parameters.

Methods

We identified 13 publications (including 13 articles and 6 supplementary materials) in patients with Atopic Dermatitis (AD), identified from standard medical databases (e.g. PUBMED, MEDLINE). A vector database was used to store the publications. The giles® AI system architecture accelerated the data extraction and SLR process through a large language model (LLM) under the Google Gemini 2.5 Pro model. Extraction from multiple publications was undertaken in parallel, enhancing the speed of extraction.

From a total of 13 publications, the following parameters were identified for extraction:

- Treatment arm
- Therapy type (monotherapy or combination therapy)
- Population (entire study population or reported subgroups)
- Age group (children, adolescents or adults)
- Disease severity (mild, moderate or severe)
- Assessment timepoint (week of assessment)
- Period of treatment (early, induction, maintenance or follow-up)
- Eczema Area and Severity Index (EASI) percentage change from baseline: including n, mean, standard deviation (SD), standard error (SE), lower and upper 95% confidence intervals
- EASI50, EASI75, EASI90 (numerators, denominators and % of people achieving 50%, 75% and 90% improvement in EASI from baseline)
- Investigator's Global Assessment (IGA) of 0 – clear or 1 – almost clear (numerators, denominators and % of people achieving IGA 0 or 1)



.PDF
.DOCX

Vector DB

19 scientific documents regarding Atopic Dermatitis



Data Extraction performed by giles®

using Google Gemini 2.5 Pro model



2118 data points extracted

1266 had reported data

Evaluation Process

Out of the 13 publications, two were used for validation during the development of the data extraction algorithm and prompt engineering.

The algorithm within giles® receives as input a Comma-Separated Values (CSV) file containing a row of pre-populated key headings. These headings are then populated with the relevant data from the publications, uploaded to the giles® AI research agent.

For each publication uploaded into giles®, the algorithm outputs multiple rows containing either the key data required or "NR" ("Not Reported") for any table cells for which data were not found or for which the LLM was unable to locate data. Given that the key data to be extracted is described through a CSV at run-time, this maintains the algorithm and LLM agnostic of the medical subject of the publications.

Metrics

Data extraction was performed in two steps. In the first step, three iterations of data extraction were performed on the same two Atopic Dermatitis publications used for validation during the development of the algorithm. During these three iterations, the table headings within the extraction sheet were manually refined in order to maximize the correctly extracted data. A cell in the extracted table is assumed to be correct if:

- a data point is present in the publication, and the exact value is extracted, or
- a data point is not present in the publication, and the cell is filled out with "NR".

In the second step, the refined data extraction sheet was used to run the final data extraction on the remaining 11 publications.

The following metrics were used to determine the performance of extraction:

- Overall accuracy – how often was a cell of the extraction table populated correctly:

$$\text{Overall Accuracy} = \frac{Correctly\ populated\ cells}{Total\ populated\ cells} \quad (1)$$

- Reported data accuracy – from all the times that data were reported, how often did the algorithm extract the data correctly:

$$\text{Reported Accuracy} = \frac{Correct\ Value}{Total\ Reported\ Data\ Points} \quad (2)$$

- NR Precision - from all the times that the algorithm filled a cell with NR, how often were the data NR in the publication. Correctly assigned NR cells are considered true positives and incorrectly assigned NR cells are considered false positives:

$$\text{NR Precision} = \frac{true\ positives}{true\ positives + false\ positives} \quad (3)$$

- NR Recall - from all the times that the data should have been NR, how often did the algorithm say NR. Cells that have hallucinated values and should have been assigned NR are considered false negatives:

$$\text{NR Recall} = \frac{true\ positives}{true\ positives + false\ negatives} \quad (4)$$

Results

From a total of 2118 data points that were extracted, 1266 (60%) had available data and 852 (40%) were not reported (NR). Table 1 summarizes the average values of the analyzed metrics, computed across the articles (11) and supplementary materials (6) from the remaining 11 publications.

Overall Accuracy	Reported Data Accuracy	NR Precision	NR Recall
0.95	0.94	0.95	0.98

Table 1. Average metrics calculated across 11 articles and 6 supplementary materials – a total of 17 sources.

Conclusion

The giles® AI research agent for SLR extraction had high accuracy approaching 95% for all measures (overall accuracy, reported data accuracy, precision and recall). In particular low hallucination was reported: when data were not reported in a paper, giles® provided a data point only 2% (1 − NR Recall) of the time.

This has shown giles® to be an exceedingly versatile and accurate tool for SLR which would speed up the manual intensive and time-consuming data extraction in the SLR creation. This could have a major impact on delivery timelines of Health Technology Assessment (HTA) reporting.

Future work

For future work we aim to:

- Implement per-cell references for grounding and fast verification of the algorithm's outputs.
- Explore automatic suggestion of data points to be extracted given a set of publications.
- Automatic identification of the publications while already being connected to databases.
- Connect to regulatory/HTA databases, and proprietary IP.

[1] G. Gartlehner., L. Kahwati, R. Hilscher, et al., "Data extraction for evidence synthesis using a large language model: A proof-of-concept study", Res Syn Meth, vol. 15, no. 4, pp. 576-589, 2024, doi:10.1002/jrsm.1710;
[2] L. Schmidt, K. Hair, S. Graziozi, F. Campbell, C. Kapp, A. Khanteymoori, D. Craig, M. Engelbert, J. Thomas, "Exploring the use of a Large Language Model for data extraction in systematic reviews: a rapid feasibility study", 2024. 10.48550/arXiv.2405.14445.
[3] T. Helms Andersen, T.M. Marcussen, A.D. Termannsen, T.W.H. Lawaetz, O. Nørgaard, "Using Artificial Intelligence Tools as Second Reviewers for Data Extraction in Systematic Reviews: A Performance Comparison of Two AI Tools Against Human Reviewers", Cochrane Evid Synth Methods, vol. 3, no. 4, July 2025. doi: 10.1002/cesm.70036; PMID: 40661122; PMCID: PMC12257877.