SLA-067

A FULLY AUTOMATED AI-ENABLED PIPELINE FOR EXTRACTING KAPLAN-MEIER SURVIVAL DATA FROM SCIENTIFIC PLOTS

+LCP powering possibility

▶ WATCH THE AI WALKTHROUGH
🎧 VIDEO + PODCAST

Open your camera app & scan
No special software required

Zdorovtsova N[1], Chung R[1], Tsvetanova A[1], Bray B*[1]
*Corresponding author: Dr Ben Bray (ben.bray@lcp.uk.com)

[1] Health Analytics, Lane Clark & Peacock LLP, London, UK

Summary

+ The time that it takes to extract quantitative data from plots continues to be a key barrier to rapid research synthesis.

+ We developed and validated an AI-enabled, automatic data extraction pipeline to extract data from Kaplan-Meier plots published in scientific publications, and re-estimated the source data using statistical methods.

+ The accuracy of our automated data extraction pipeline approaches that of manual digitisation. The key remaining challenges are handling plots with unusual formatting. While efforts continue to improve the robustness of our method, automated extraction shows promise in expediting evidence synthesis at scale.

Background +

- The ability to extract quantitative data from scientific plots at scale unlocks new possibilities for evidence synthesis and real-world evidence research, particularly when original datasets are unavailable.
- In clinical research, Kaplan-Meier survival curves are widely used to present time-to-event data.
- However, existing digitisation methods often require substantial manual effort, limiting their scalability and reproducibility[1,3].

Objectives +

To **develop and validate a computer vision pipeline that automates plot digitisation**, enabling large-scale extraction of survival data with minimal human intervention.

Methods +

We developed and validated an automated computer vision pipeline using **convolutional neural networks, vision-enabled large language models, and edge detection methods**. The validation dataset comprised published Kaplan-Meier plots from peer-reviewed scientific literature published between 2007-2024. Real-world plots were manually digitised using the online tool WebPlotDigitzer and were used as the baseline. Primary and secondary outcomes are reported in Figure 1.

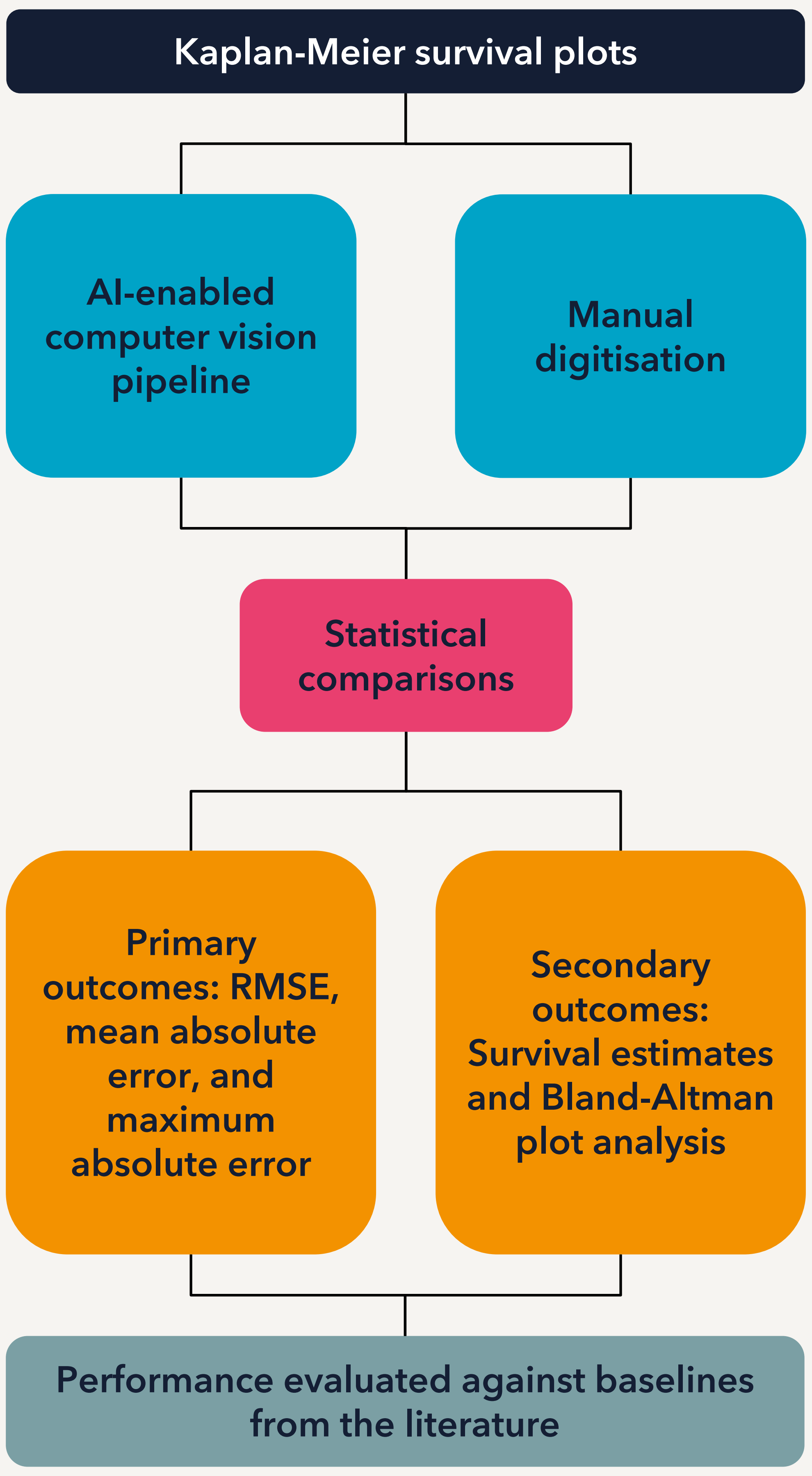
Errors between the manual and automated approach were calculated by taking the difference in survival probability at each shared time value across both curves.

Performance was evaluated against pre-specified criteria from literature: RMSE ≤0.05, mean absolute error ≤0.02, maximum absolute error ≤0.05[3], and median survival time within 10% of published estimates. Metrics were adjusted for comparability across different plot scales and formats.

Results +

- 29 curves across 9 real-world plots were evaluated and digitised both manually and using our automated pipeline.
- Between manually and automatically digitised coordinates, average RMSE was 0.13 (SD [0.06]) average mean absolute error was 0.11 (SD [0.06]), and average maximum absolute error was 0.28 (SD [0.12]).
- The average median survival time from manually digitised curves diverged -0.19% from the published median survival time, compared to 1.86% when using our automated pipeline.
- Though accuracy was lower for automatically extracted coordinates, compared to manually extracted coordinates, our progress represents a significant engineering step towards automation (see Figure 2).

Figure 1: Methodology schematic



Kaplan-Meier survival plots

AI-enabled computer vision pipeline

Manual digitisation

Statistical comparisons

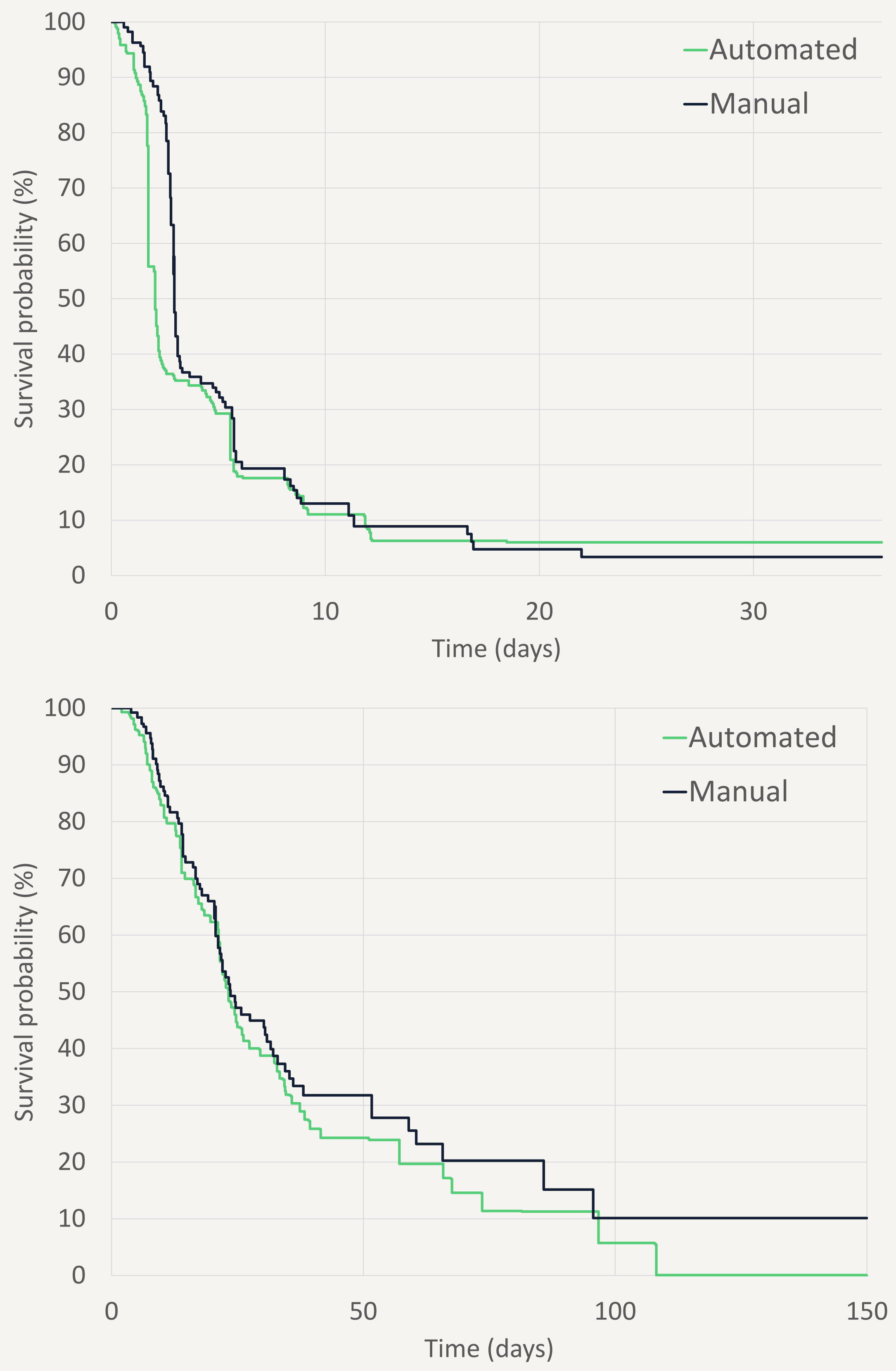
Primary outcomes: RMSE, mean absolute error, and maximum absolute error

Secondary outcomes: Survival estimates and Bland-Altman plot analysis

Performance evaluated against baselines from the literature

Figure 2: Examples of extracted curves

Manually-extracted (black) versus automatically-extracted (green) coordinates from two real-world Kaplan-Meier plots[4,2].



Conclusions +

Our automated digitisation pipeline demonstrates **accuracy approaching that of manual extraction while significantly reducing human effort**. The tool's performance across diverse plot styles supports its utility for large-scale data extraction in real-world evidence research, though further validation is needed for heavily annotated or non-standard figures. These results suggest potential for improving reproducibility and efficiency in pharmacoepidemiological research.

References:
1. Guyot et al. (2012). Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. DOI: 10.1186/1471-2288-12-9
2. Hassel et al. (2023). Three-Year Overall Survival with Tebentafusp in Metastatic Uveal Melanoma. DOI: 10.1056/NEJMoa2304753
3. Liu et al. (2021). IPDfromKM: reconstruct individual patient data from published Kaplan-Meier survival curves. DOI: 10.1186/s12874-021-01308-8
4. Yan et al. (2021). A Comparison of Hypofractionated and Twice-Daily Thoracic Irradiation in Limited-Stage Small-Cell Lung Cancer: An Overlap-Weighted Analysis. DOI: 10.3390/cancers13122895

Disclosure statement: The authors have nothing to disclose in relation to this research.