EpiCurveBench: Evaluating epidemic curve digitization

Thomas Berkane* and **Maimuna S. Majumder**
Computational Health Informatics Program, Boston Children's Hospital & Harvard Medical School, Boston, MA 02115, USA

Abstract

Accurate data on disease case counts over time is essential for training reliable disease forecasting models. However, such data is often locked in non-machine-readable formats, most commonly as epidemic curve (epicurve) images—charts that depict case counts over time for a given location. Digitizing these charts would greatly expand the data available for forecasting models, improving their accuracy. Manual digitization, though, is very time-consuming, and existing automated methods struggle with real-world epicurves due to dense data points, overlapping series, and diverse visual styles. To address this, we present EpiCurveBench, a benchmark of 100 manually curated and annotated epicurve images collected from diverse sources. The dataset spans a wide range of chart styles, from simple to highly complex. We also introduce a new evaluation metric, based on Edit Distance with Real Penalty, that captures the temporal structure of epicurves, handles series of varying lengths, and remains stable in the presence of incomplete data. Using this metric, we evaluate state-of-the-art chart data extraction methods on EpiCurveBench and find substantial room for improvement, with the best model achieving only 25%. We release the dataset and evaluation pipeline to accelerate progress in epicurve extraction. More broadly, EpiCurveBench provides a challenging testbed to advance research in chart data extraction beyond epidemic forecasting.

Keywords: Dataset, Vision-Language Models, Chart Data Extraction, Epidemiology

1. Introduction

Models that forecast disease case counts over time can reduce disease burden (Brooks et al., 2015), but they require historical temporal case data, often sourced from disease outbreak reports. Much of this data exists only as epidemic curve images (epicurves)—charts showing case counts over time for a location—and is thus effectively "locked away." Epicurves are commonly produced by ministries of health in endemic regions or during outbreaks, but their non-machine-readable format limits their use for training forecasting models. Access to the underlying data would also enable analyses of historical transmission patterns, vaccine effectiveness, and global resource allocation. During outbreaks, public health officials are often asked to compile and format such data on demand—a burdensome, inefficient process—highlighting the need for scalable methods to automatically extract and structure epicurve data.

Currently, manual tools like WebPlotDigitizer (Rohatgi) are used to digitize epicurves, but extracting data from a single complex image can take hours. Automated chart data extraction aims to recover the underlying numerical data from chart images. Early methods relied on traditional computer vision and optical character recognition (Farahani et al., 2023). Recently, large vision-language models have achieved stronger performance (Han et al., 2023; Xia et al., 2025). However, these models still struggle with complex, real-world charts. Most existing datasets (Methani et al., 2020; Masry et al., 2022; Xia et al., 2025) consist of synthetic charts with low diversity. A notable exception, ChartQA (Masry et al., 2022), includes real-world charts, but they come from only four platforms and are generally simple—basic infographics, pie charts, or bar charts with few data points and clearly printed numbers. The lack of diverse, challenging benchmarks has limited progress on more robust extraction methods.

To address this, we introduce EpiCurveBench, a benchmark designed to advance automated epicurve data extraction and benefit the broader chart data extraction community. EpiCurveBench contains 100 carefully curated, hand-annotated epicurve images from diverse sources, spanning a wide range of coun-

* Correspondence: thomas.berkane@childrens.harvard.edu

© T. Berkane & M.S. Majumder.

tries, years, and styles. It includes 30 simpler epicurves for a smooth transition into the task and 70 challenging ones selected for diversity and difficulty. These test a model's ability to generalize, with many featuring unique visual formats. Extracting data from them requires fine-grained visual understanding and adaptability, as epicurves often include dense data points, overlapping series, rotated or tiny text, low-resolution images, thin bars, low contrast, and overlaid text. Even human annotators found the task difficult: some epicurves required up to three hours to digitize using tools like WebPlotDigitizer.

Existing evaluation metrics for chart data extraction are poorly suited for epicurves, as they ignore temporal structure, overpenalize small shifts, and fail to distinguish missing from incorrect values. To address this, we introduce EpiCurve Similarity (ECS), a measure based on Edit Distance with Real Penalty (Chen and Ng, 2004), which aligns two time series through minimal-cost matching while explicitly handling gaps and temporal misalignments, critical for accurate epicurve evaluation.

We evaluate multiple state-of-the-art automated chart data extraction methods, including frontier vision-language models, on EpiCurveBench using this new metric. Results show that EpiCurveBench is highly challenging: current state-of-the-art methods leave substantial room for improvement, unlike other chart benchmarks that are nearing saturation. EpiCurveBench highlights significant gaps in current methods and lays the groundwork for improving both epidemic forecasting and chart data extraction, advancing visual reasoning in foundation models.

Contributions. (1) EpiCurveBench, a challenging, real-world benchmark for epicurve data extraction. (2) ECS, a metric tailored to temporal curve evaluation. (3) A comprehensive evaluation of state-of-the-art methods. (4) Open-source evaluation code to support development of new methods.

2. EpiCurveBench

2.1. Data collection

We collect epicurve images from diverse sources to capture a wide range of styles and formats. We crawl directory pages from organizations such as WHO regional offices, PAHO, and the CDC, scraping images from webpages and linked outbreak report PDFs. We also run targeted Google Image searches with varied query terms to capture additional images.

This process yields 1,397 epicurve images, from which we manually curate a final set of 100 that are the most diverse and challenging. These are split into two subsets: 30 Basic epicurves, with no more than two series per chart and series lengths of 30 points or fewer, and 70 Advanced epicurves selected for maximum stylistic diversity and difficulty. The Basic subset provides an easier entry point for developing and testing extraction methods.

Appendix A shows the distribution of sources for the curated dataset.

2.2. Data annotation

We manually annotate all epicurve images using WebPlotDigitizer, which supports precise point-and-click data extraction. The process begins by calibrating the x- and y-axis scales using two known points on each axis. Once calibrated, the annotator clicks each data point on the chart to record its x and y values.

Each epicurve produces a CSV file containing the extracted data. The first column lists x-axis values, followed by one or more columns for the y-axis values of each series. We use the original axis labels and series names in the CSV header for clarity and consistency. When multiple series are present, we extract all of them, including non-case-count data such as rainfall or other covariates. Appendix B gives details on handling ambiguous annotation cases, which are also provided to the extraction methods to ensure consistency between human and automated outputs.

A single annotator completes the initial annotations. To assess reliability, a second annotator independently annotates 20 randomly selected curves from the 70 Advanced epicurves, yielding an inter-annotator ECS of 93.78% (see Section 2.4).

Annotation times range from 2 to 179 minutes, with a mean of 23.7 minutes, a median of 16 minutes, and a standard deviation of 24.7 minutes.

2.3. Dataset characteristics

Figure 1 shows sample epicurves. The first row contains three examples from the Basic subset, while the remaining rows show Advanced epicurves, illustrating diverse styles and formats. Challenges include thin bars, low resolution, heavy text overlays, stacked bars with similar colors, historical plots with hard-to-read values, and multi-panel sharing an x-axis.

The benchmark includes 54 bar charts, 14 line charts, and 32 mixed charts. Series per image range from 1 to 14 (mean 2.8, median 2, SD 2.5). Series

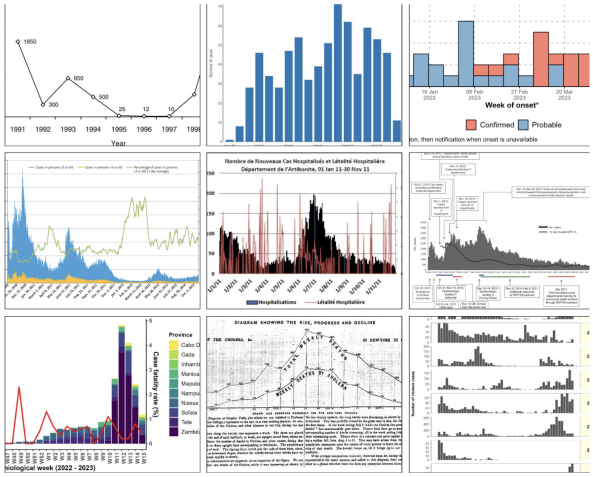

Figure 1: Sample images from EpiCurveBench. Parts of the images, including axes, are truncated for space.

lengths range from 5 to 715 points (mean 89, median 58, SD 122.8). Charts span 31 countries and the years 1849–2025. Image widths range from 334–1650 pixels and heights from 148–1434 pixels.

This diversity in chart types, time periods, and visual styles makes EpiCurveBench especially challenging, requiring models to generalize across highly variable conditions for robust performance.

2.4. Metric

Two metrics are commonly used for evaluating chart data extraction : Relative Mapping Similarity (Liu et al., 2023, RMS) and Structuring Chart-oriented Representation Metric (Xia et al., 2024, SCRM). Both are poorly suited to epicurves.

First, they treat extracted points as unordered sets, ignoring temporal structure. As a result, a curve shifted by a few days is penalized as harshly as a completely incorrect one. Many extraction errors are temporal, such as missing early data or slight misalignments. Second, they fail to distinguish between missing points and incorrect values, treating absent points as complete mismatches. Finally, they overemphasize textual elements like axis labels, which are often sparse in epicurves, while numerical time series values are more important for downstream modeling.

Epicurve evaluation requires handling temporal misalignments, differing sequence lengths, and varying sampling frequencies. For example, ground truth may be reported daily while extractions are weekly, or contain spurious points or truncated ends. Standard pointwise metrics assume perfect alignment and equal length, which is unrealistic.

To address these issues, we introduce EpiCurve Similarity (ECS), a new metric specifically designed for epicurve extraction. ECS builds on Edit Distance with Real Penalty (Chen and Ng, 2004, ERP), a dynamic programming method that explicitly handles gaps and unmatched regions. ERP extends classical edit distance by using real-valued penalties for differences between matched points instead of fixed penalties. It supports insertions, deletions, and substitutions while constraining the temporal range of matches using a Sakoe–Chiba band (Sakoe and Chiba, 1978), which we set to seven time steps (one week).

Let $E = \{e_1, \ldots, e_m\}$ be the set of series extracted from an image and $T = \{t_1, \ldots, t_n\}$ the corresponding set of ground-truth series. Since epicurve images often contain multiple series, we first match each extracted series to its ground-truth counterpart using Normalized Levenshtein Distance on series labels (e.g., from legends or y-axis labels), following Liu et al. (2023); Furkan Biten et al. (2019). Unmatched extracted series are ignored, and unmatched ground-truth series are scored as zero.

For a matched pair $p = (p_1, \ldots, p_P)$ and $t = (t_1, \ldots, t_T)$, ERP is computed using the recurrence relation:

$$C(i,j) = \min \begin{cases} C(i-1,j-1) + D_\theta(p_i, t_j) \\ C(i-1,j) + \lambda \\ C(i,j-1) + \lambda \end{cases}$$

with $C(0,0) = 0$, $C(i,0) = i \cdot \lambda$, and $C(0,j) = j \cdot \lambda$. We fix the gap penalty at $\lambda = 1$ (as in Liu et al. (2023)), and the recurrence is evaluated only for $|i - j| \leq \sigma$, where $\sigma = 7$, limiting matches to within a one-week window.

The distance between two points is calculated using a relative distance function:

$$D_\theta(p_i, t_j) = \min\left(1, \frac{|p_i - t_j|}{|t_j|}\right),$$

where differences below $\theta = 0.1$ are penalized proportionally, and larger differences are treated as maximally dissimilar, following Liu et al. (2023). The ERP distance is then:

$$\mathrm{ERP}(p,t) = C(P,T),$$

representing the minimum cost of transforming p into t.

ECS normalizes this cost to produce a similarity score between 0 and 1:

$$\mathrm{ECS}(p,t) = 1 - \frac{\mathrm{ERP}(p,t)}{\#\mathrm{matches} + \#\mathrm{gaps}}.$$

For each image, ECS averages similarity scores across matched series, assigning zero to unmatched ground truth series, and then computes the dataset-level average.

3. Baseline performance

3.1. Methods evaluated

We evaluate several methods on EpiCurveBench, including two state-of-the-art vision-language models specialized in chart data extraction: OneChart (Chen et al., 2024) and TinyChart (Zhang et al., 2024). We also test GPT-5 under four configurations: minimal, medium, and high reasoning effort, as well as high reasoning effort with access to a code interpreter tool. For GPT-5, we use the prompt in Appendix C and a temperature of 0.

3.2. Results

Table 1 reports ECS scores for both subsets and the total inference costs for processing the 100 images.

Table 1: Performance of all evaluated methods on EpiCurveBench. Results are reported as ECS scores (%) for the Basic and Advanced subsets, along with total inference costs (USD) for the 100 images.

Model	Basic (%)	Advanced (%)	Cost
OneChart	39.48	5.57	–
TinyChart	39.77	8.74	–
GPT-5 Minimal	58.80	**25.25**	$0.71
GPT-5 Medium	57.12	17.15	$3.40
GPT-5 High	62.50	19.73	$6.81
GPT-5 High+Code	**63.89**	21.00	$5.51

The results show that EpiCurveBench is challenging for all tested approaches and far from saturation. The best-performing models reach an ECS of only 63.89% on the Basic subset and 25.25% on the Advanced subset.

OneChart and TinyChart lag behind GPT-5 on both subsets. Their limited context windows cannot accommodate the many data points in complex epicurves, leading to truncation. Future work could explore extracting series individually to mitigate this. OneChart often misreads series and axis labels, preventing correct matching with ground truth and creating a performance ceiling. TinyChart handles labels better, likely due to its larger parameter count, but still produces highly inaccurate values.

Among GPT-5 configurations, the minimal reasoning setting performs surprisingly well, slightly outperforming medium on the Basic subset and achieving the highest performance overall on the Advanced subset. This suggests that minimal reasoning avoids errors sometimes introduced by additional reasoning steps. Adding the code interpreter tool slightly improves performance while reducing inference cost.

Across GPT-5 configurations, common errors persist. While generally strong at reading labels, minimal and medium settings sometimes miss entire series. Higher reasoning helps but still struggles with precise extraction from densely packed points. These challenges reflect a broader limitation of multimodal reasoning models: difficulty distinguishing fine-grained visual elements (Liu et al., 2025), especially in charts (Razeghi et al., 2024). Extracting dense time series thus remains an open problem.

4. Limitations

EpiCurveBench includes only 100 annotated images, which may not fully represent the diversity of global epidemic curve formats. All annotations are created by a single annotator, introducing potential labeling noise. Model evaluations use basic prompting without fine-tuning, so performance may underestimate what could be achieved with further tuning.

5. Conclusion

We introduce EpiCurveBench, a challenging benchmark for epidemic curve extraction, along with a new metric and open-source evaluation pipeline. Our evaluation of state-of-the-art models reveals a significant gap between current methods and the complexity of real-world epicurves. Future work will explore advanced reasoning and code generation approaches that adapt to diverse chart styles and formats. We aim to drive the development of generalizable extraction methods to support better epidemic forecasting.

References

Logan C. Brooks, David C. Farrow, Sangwon Hyun, Ryan J. Tibshirani, and Roni Rosenfeld. Flexible modeling of epidemics with an empirical bayes framework. *PLOS Computational Biology*, 11(8): 1–18, 08 2015. doi: 10.1371/journal.pcbi.1004382. URL https://doi.org/10.1371/journal.pcbi.1004382.

Jinyue Chen, Lingyu Kong, Haoran Wei, Chenglong Liu, Zheng Ge, Liang Zhao, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Onechart: Purify the chart structural extraction via one auxiliary token, 2024. URL https://arxiv.org/abs/2404.09987.

Lei Chen and Raymond Ng. On the marriage of lp-norms and edit distance. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30*, VLDB '04, page 792–803. VLDB Endowment, 2004. ISBN 0120884690.

Ali Farahani, Peyman Adibi, Alireza Darvishy, Saeed Ehsani, and Hans-Peter Hutter. Automatic chart understanding: A review. *IEEE Access*, PP:1–1, 01 2023. doi: 10.1109/ACCESS.2023.3298050.

Ali Furkan Biten, Rubèn Tito, Andrés Mafla, Lluis Gomez, Marçal Rusiñol, Minesh Mathew, C.V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Icdar 2019 competition on scene text visual question answering. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1563–1570, 2019. doi: 10.1109/ICDAR.2019.00251.

Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. Chartllama: A multimodal llm for chart understanding and generation, 2023. URL https://arxiv.org/abs/2311.16483.

Chengzhi Liu, Zhongxing Xu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models, 2025. URL https://arxiv.org/abs/2505.21523.

Fangyu Liu, Julian Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. DePlot: One-shot visual language reasoning by plot-to-table translation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10381–10399, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.660. URL https://aclanthology.org/2023.findings-acl.660/.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning, 2022. URL https://arxiv.org/abs/2203.10244.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots, 2020. URL https://arxiv.org/abs/1909.00997.

Yasaman Razeghi, Ishita Dasgupta, Fangyu Liu, Vinay Venkatesh Ramasesh, and Sameer Singh. Plot twist: Multimodal models don't comprehend simple chart details. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5922–5937, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.342. URL https://aclanthology.org/2024.findings-emnlp.342/.

Ankit Rohatgi. Webplotdigitizer. URL https://automeris.io.

H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978. doi: 10.1109/TASSP.1978.1163055.

Renqiu Xia, Haoyang Peng, Hancheng Ye, Mingsheng Li, Xiangchao Yan, Peng Ye, Botian Shi, Yu Qiao, Junchi Yan, and Bo Zhang. Structchart: On the schema, metric, and augmentation for visual chart understanding, 2024. URL https://arxiv.org/abs/2309.11268.

Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Peng Ye, Min Dou, Botian Shi, Junchi Yan, and Yu Qiao. Chartx chartvlm: A versatile benchmark and foundation model for complicated chart reasoning, 2025. URL https://arxiv.org/abs/2402.12185.

Liang Zhang, Anwen Hu, Haiyang Xu, Ming Yan, Yichen Xu, Qin Jin, Ji Zhang, and Fei Huang. TinyChart: Efficient chart understanding with program-of-thoughts learning and visual token merging. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1898, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.112. URL https://aclanthology.org/2024.emnlp-main.112/.

Appendix A. Distribution of sources

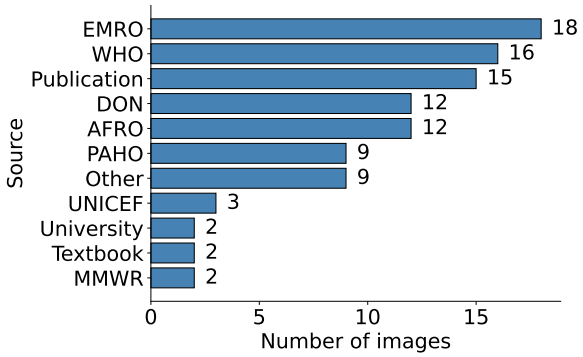

Figure 2: Distribution of EpiCurveBench sources by origin. There are 19 unique source types. The "Other" category includes sources that appear only once. Most publication sources come from platforms such as ScienceDirect and ResearchGate.

Appendix B. Handling Annotation Ambiguities

- Only data that is visually present and clearly labeled is annotated. No assumptions are made about missing or implied data.
- Each visible segment in a stacked bar chart is annotated as a separate time series if the figure includes distinct segment labels in the legend or directly on the chart.
- If segment labels are missing, incomplete, or visually indistinguishable, only the aggregate total (height of the full stack) is annotated.
- In multi-panel figures, series in different panels are annotated as being part of the same chart.

Appendix C. GPT-5 Prompt

> Here is an image of a chart. Please extract the numerical data it represents and return it in CSV format with appropriate headers. Copy the headers exactly as they are in the image. Make sure to extract data for all x axis values, even those not written on the image, without leaving any gaps. Remember: The sole output should be the CSV table. Nothing else.