Supplemental Material

Data S1.

Supplemental Methods

Natural Language Processing

Blocks of fully de-identified text from each Echo Report were parsed into a CSV, containing a row for each report section indicating the study id, report section title, and text for that section. This data was then processed by an NLP engine built by Echo IQ Ltd (www.echoiq.ai) to determine a label for each of the echo studies indicating the physician reported individual wall motion score, including the interpretation of each of the 17 myocardial segments and specific severity allocation (normal, mild hypokinesis, severe hypokinesis, akinesis, dyskinesis and aneurysm). The NLP engine processes data across a number of stages: 1) **Pre-processing** to fix misspellings and formatting issues. 2) **Text extraction** to identify echo-related content. 3) **Quantitative extraction** to extract measurements. 4) **Qualitative extraction** to extract specific reference to the presence of and disease grading reported in the body of the report. This includes the presence or absence of a wall motion abnormality, accompanying comments on wall motion severity, and any additional comments. All relevant comment terms were included in the proprietary NLP system (for example, location of the wall motion abnormality, infarction, aneurysm etc). Chronicity was established using specific search terms. The presence of any prior valve intervention was extracted, and these patients were excluded from analysis.

The NLP engine uses a range of techniques including regular expression parsing, open-source NLP libraries and heuristic rules, and were also combined with specialist-informed custom dictionaries for cardiovascular terms and misnomers (e.g., 'anormality' replaced with 'abnormality). Performance of the heuristic rules and identified dictionary values was tuned by processing text across the entire corpus of echo text in NEDA. Performance and cross checking of the NLP engine was performed using Boolean and string extractions, and random manual comparisons of edge cases.

Figure S1. The Echo IQ NLP engine, with some examples of the outputs obtained from text extraction.