Supplementary information

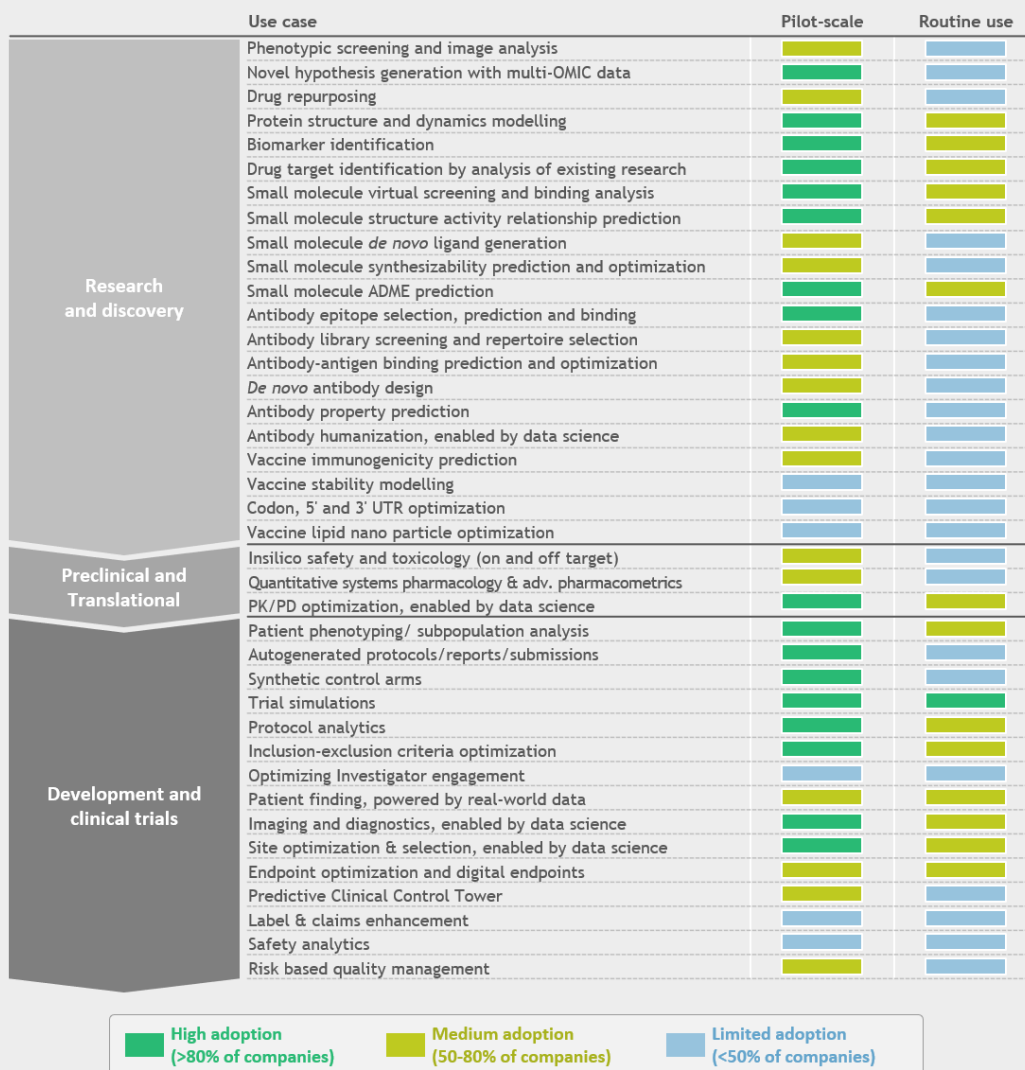# Data science in pharmaceutical R&D: the DISRUPT-DS industry roundtable

In the format provided by the authors and unedited
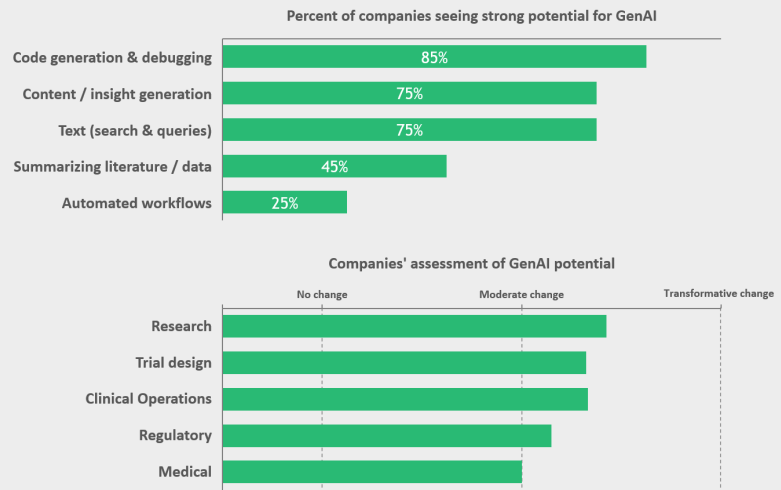
Benchmarking methodology

Companies participating in the DISRUPT-DS roundtable were benchmarked by Boston Consulting Group in mid-2023, and aspects of the benchmarking (specifically Supplementary Figure S1) were refreshed in mid-2024. Benchmarking involved a series of structured interviews and, for the refresh, a survey of data science stakeholders across research/discovery, preclinical/translational and development. With each company participating in DISRUPT-DS, between 3 and 5 interviews were conducted, each of which lasted approximately one hour. The results were subsequently processed for consistency, analyzed, anonymized and aggregated. Please note: The roundtable participants collectively decided not to include CMC in the benchmarking.
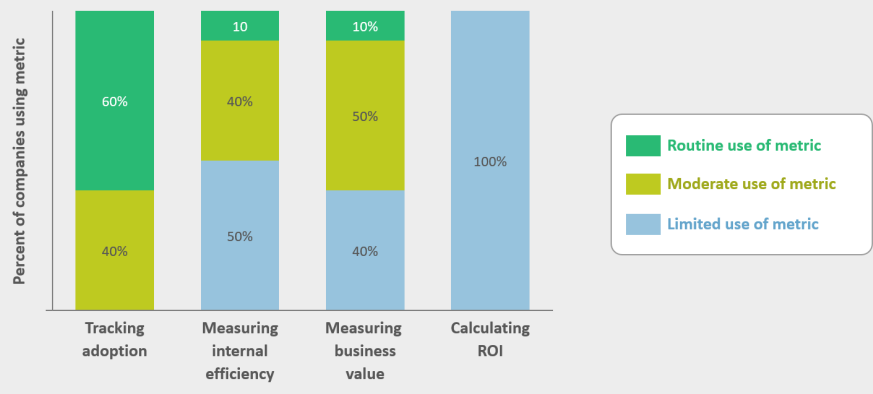
Key results of benchmarking exercise

Supplementary Figure S1 | Adoption of data science use cases across the R&D value chain by companies participating in the DISRUPT-DS roundtable. The list of use cases was defined through interviews of roundtable participants. 'Pilot-scale' means that use cases are applied to a limited number of R&D programs. 'Routine use' means that use cases are applied regularly across R&D programs in one or more therapy areas. Please note: The benchmarking does not include CMC use cases.

Stage	Use case	Pilot-scale	Routine use
Research and discovery	Phenotypic screening and image analysis	Medium	Limited
	Novel hypothesis generation with multi-OMIC data	High	Limited
	Drug repurposing	Medium	Limited
	Protein structure and dynamics modelling	High	Medium
	Biomarker identification	High	Medium
	Drug target identification by analysis of existing research	High	Medium
	Small molecule virtual screening and binding analysis	High	Medium
	Small molecule structure activity relationship prediction	High	Medium
	Small molecule *de novo* ligand generation	Medium	Limited
	Small molecule synthesizability prediction and optimization	Medium	Limited
	Small molecule ADME prediction	High	Medium
	Antibody epitope selection, prediction and binding	High	Limited
	Antibody library screening and repertoire selection	Medium	Limited
	Antibody-antigen binding prediction and optimization	High	Limited
	De novo antibody design	Medium	Limited
	Antibody property prediction	High	Limited
	Antibody humanization, enabled by data science	Medium	Limited
	Vaccine immunogenicity prediction	Medium	Limited
	Vaccine stability modelling	Limited	Limited
	Codon, 5' and 3' UTR optimization	Limited	Limited
	Vaccine lipid nano particle optimization	Limited	Limited
Preclinical and Translational	Insilico safety and toxicology (on and off target)	Medium	Limited
	Quantitative systems pharmacology & adv. pharmacometrics	Medium	Limited
	PK/PD optimization, enabled by data science	High	Medium
Development and clinical trials	Patient phenotyping/ subpopulation analysis	High	Medium
	Autogenerated protocols/reports/submissions	High	Limited
	Synthetic control arms	High	Limited
	Trial simulations	High	High
	Protocol analytics	High	Medium
	Inclusion-exclusion criteria optimization	High	Medium
	Optimizing Investigator engagement	Limited	Limited
	Patient finding, powered by real-world data	Medium	Medium
	Imaging and diagnostics, enabled by data science	High	Medium
	Site optimization & selection, enabled by data science	High	Medium
	Endpoint optimization and digital endpoints	Medium	Medium
	Predictive Clinical Control Tower	Limited	Limited
	Label & claims enhancement	Limited	Limited
	Safety analytics	Limited	Limited
	Risk based quality management	Medium	Limited

Legend:
- **High adoption** (>80% of companies)
- **Medium adoption** (50–80% of companies)
- **Limited adoption** (<50% of companies)

Supplementary Figure S2 | Assessment of the of applications of generative AI by companies participating in the DISRUPT-DS roundtable

Percent of companies seeing strong potential for GenAI

Code generation & debugging	85%
Content / insight generation	75%
Text (search & queries)	75%
Summarizing literature / data	45%
Automated workflows	25%

Companies' assessment of GenAI potential

No change   Moderate change   Transformative change

- Research
- Trial design
- Clinical Operations
- Regulatory
- Medical

Supplementary Figure S3 | Different metrics for assessing the impact of data science in biopharma R&D used by companies participating in the DISRUPT-DS roundtable

Percent of companies using metric

	Tracking adoption	Measuring internal efficiency	Measuring business value	Calculating ROI
Routine use of metric	60%	10	10%	
Moderate use of metric	40%	40%	50%	
Limited use of metric		50%	40%	100%

Legend:
- Routine use of metric
- Moderate use of metric
- Limited use of metric

Supplementary Figure S4 | Current status of responsible use of data science for companies which are participating in the DISRUPT-DS roundtable and for which this information was available

Percent of companies have implemented, as of 2023

Articulated strategy for data ethics and responsible use of data science and AI	90%
Processes in place to monitor and review criteria for responsible use of data science and AI	90%
Defined leadership and established escalation paths for responsible use of data science and AI	70%
Risk assessments in place for data ethics and responsible use of data science and AI	70%
Training in place to identify biases in data science and AI work	70%