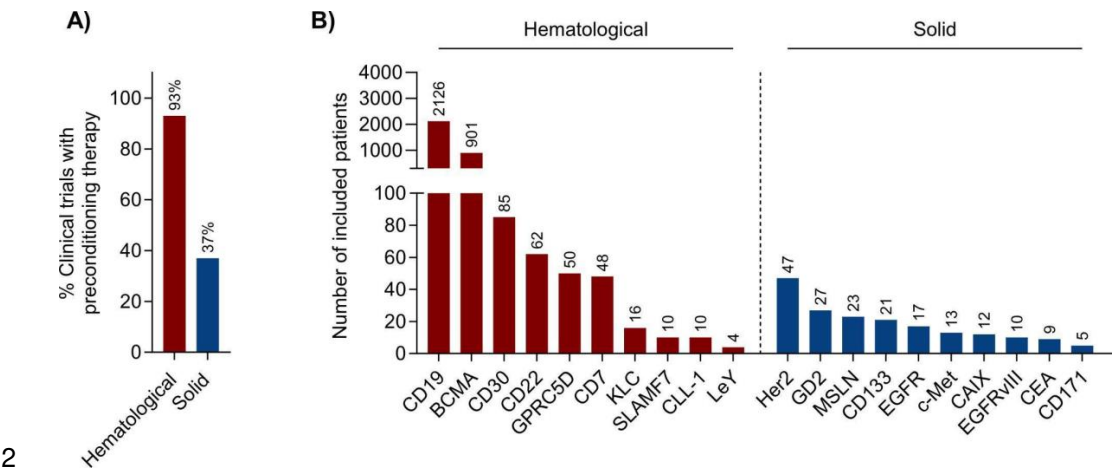


1 **Supplementary Figures**



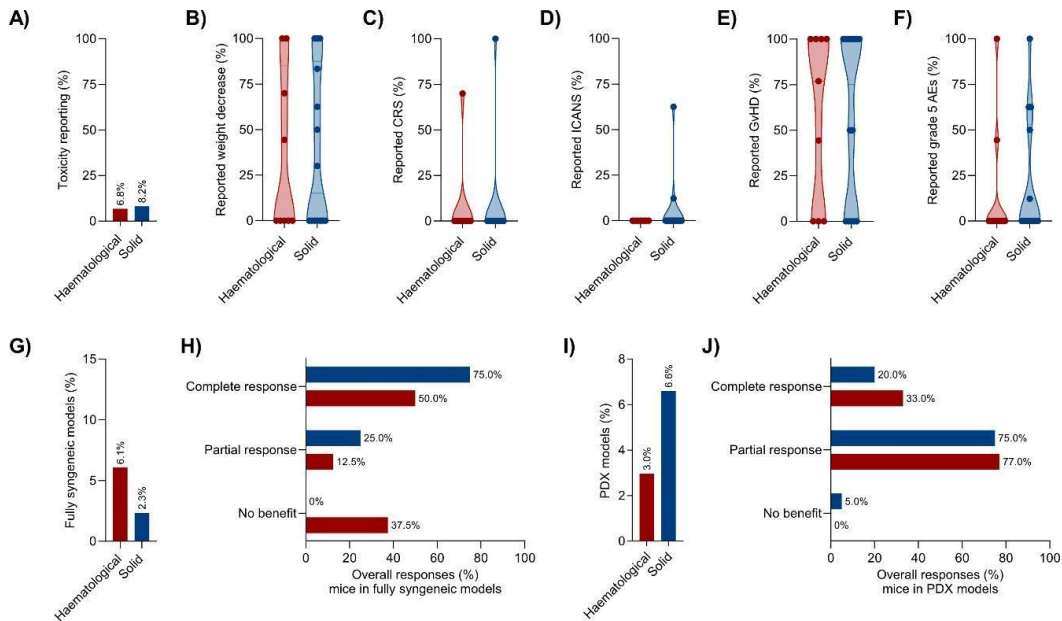
2

3 **Supplementary Figure 1: Safety data and treatment responses for**

4 **hematological and solid clinical trials of CAR-T cells.** (A) Implementation of

5 preconditioning therapy, in terms of chemo- or radiation therapy. (B) Distribution of

6 studied patients by target antigen.



7

8 **Supplementary Figure 2: Safety evaluation and treatment responses for**

9 **hematological and solid preclinical studies of CAR-T cells.** (A) Rates of

10 toxicity reporting for CAR-T cell therapy in hematological and solid tumor models.

11 (B) Proportions of toxic events in experimental animals such as weight loss (B),

12 CRS (C), ICANS (D), GvHD (E) and lethal AE (F) for either hematological or solid

13 tumors. (G) Sum of all immunocompetent mice employed in fully syngeneic tumor

14 models belonging to CAR treatment groups for either category of tumors. (H)

15 Overall responses as for tumor clearance, partial response (decreased and slower

16 growth) or no benefit for fully syngeneic mouse models. (I) Sum of all PDX models

17 implemented in preclinical studies. (J) Response distribution across all PDX

18 models used. Overall responses were classified as tumor clearance, partial

19 response (tumor decrease and slower growth) or no benefit.

20 **Supplementary Methods**

21 **Information sources, search strategy and data collection process**

22 The clinical trial records were sourced from PubMed and ClinicalTrials.gov until
23 December 1st, 2023, employing specific search criteria. The results from PubMed
24 were retrieved using the query "((CAR-T) OR (Chimeric antigen receptor)) AND
25 (Clinical Trial[Publication Type]) AND (English[Language])". Similarly, in
26 ClinicalTrials.gov, the search query "(CAR-T) OR (Chimeric antigen receptor)" was
27 used in the "other terms" search field, with an additional filter for completed and
28 terminated trials to allow for the selection of studies with reported results. To
29 streamline the dataset, papers obtained from PubMed underwent manual
30 screening to extract the clinical trial identifiers. For the clinical trial identifiers with
31 NCT format, the records were retrieved from ClinicalTrials.gov and duplicates
32 were removed. For the clinical trials with other formats, the information was
33 manually retrieved from the relevant sources, if available.

34 The preclinical records were obtained from PubMed until December 1st, 2023,
35 using the following search: "((CAR-T) OR (chimeric antigen receptor)) AND
36 ((CD19) OR (TNFRSF17) OR (CD269) OR (BCMA) OR (Siglec-2) OR (CD22) OR
37 (Mesothelin) OR (Her2) OR (ERBB2) OR (CD30) OR (TNFRSF8) OR (MSLN) OR
38 (GD2) OR (CD7) OR (ERBB1) OR (CEA) OR (CEACAM) OR (SLAMF7) OR
39 (CD319) OR (CS1) OR (LeY) OR (KLC) OR (EGFRvIII) OR (EGFR) OR (CLL-1)
40 OR (CE7) OR (L1CAM) OR (CD171) OR (AC133) OR (CD133) OR (CA9) OR
41 (CAIX) OR (HGFR) OR (c-Met)) NOT (Review[Publication Type]) NOT (Clinical
42 trial[Publication Type]) NOT (Systematic Review[Publication Type]) NOT (Meta-
43 Analysis[Publication Type]) AND (English[Language])". These included all
44 publications in English regarding CAR-T cells and the targets from the previously
45 included clinical trials (including their commonly used synonyms), excluding
46 reviews, systematic reviews, meta-analyses and studies of clinical nature.

47 To prevent biases in the assessment, each entry was evaluated in all its aspects,
48 including inclusion/exclusion criteria and data extraction, by two reviewers
49 independently. Disagreements were resolved by discussion or with the
50 intervention of a third reviewer.

51
52

Further explanation regarding eligibility criteria and selection process

An additional clarification of the exclusion criteria considered by the investigators as the most interpretative is provided: the exclusion criterion 'Combination therapy' refers to studies in which the CAR-T cell therapeutic was not the sole object of investigation for efficacy in the clinical and preclinical settings, apart from chemo- and radiotherapy. This included, among others, knock-out and/or knock-in mutations, monoclonal antibodies, bispecific CAR molecules and chimeric switch receptors.

Studies were classified as 'Not about CAR efficacy' in the event of studies investigating, for example, the safety profile of certain suicide switches, novel methods of CAR generation and *ex vivo* expansion.

Studies were classified as 'Non-classical CAR setting' where the main study subject was a CAR molecule modified in its classical structure and mode of function. Examples include, but are not limited to, VHH-, nanobody-, or NKG2D-based CAR platforms, SynNotch and logic-gated systems.

Studies were excluded for 'Incomplete data' if they did not provide relevant information about the full CAR structure and/or the efficacy of the treatment.

Preclinical work deemed to be 'Studying irrelevant variables in the experimental set-up' was characterized by an effort to elucidate other aspects of the adoptive use of CAR-T cells other than efficacy in animal models, examples being naïve vs effector phenotypes, higher affinity binding moieties, mapping of intracellular phosphorylation patterns or the composition of the TME.

Machine learning-guided analysis of our database

Data collection and preprocessing

The input data for the Machine Learning analysis comprised both preclinical and clinical CAR-T cell studies. The preclinical dataset consisted of 303 data points, each representing a distinct study with features such as 'Target', 'Solid or Hematologic tumors', 'Entity category', 'Cancer source', 'Antigen origin', 'Ag OE' (Antigen Over-Expression), 'Injection site', 'Mouse type', 'CAR generation', 'scFv source', 'TM domain', 'Costimulatory domain', 'Preconditioning', and 'Number of Mice' (Extended data file 1). The response variable for the preclinical dataset was initially categorized into three classes ('No response', 'Partial response', 'Complete response') and subsequently binarized into 'No response' (negative class) and a

combined positive class comprising the other two categories to train the logistic regression models.

The clinical dataset consisted of 105 data points with features including 'Target', 'Solid or Hematologic tumors', 'Entity category', 'CAR generation', 'scFv source', 'TM domain', 'Costimulatory domain', 'Preconditioning', and 'Number of Participants' (Table 2 and 3).

Out of the initial set of features, 'Entity category' was removed, since several clinical trials include multiple entities in the same study. 'Number of Mice' and 'Number of Participants' were also neglected, as they are not expected to have an effect on the outcome of the trial or preclinical study.

The response variable ORR, originally a continuous variable ranging from 0 to 1, was binarized using a 0.25 cutoff. The 'Preconditioning' variable was also binarized to indicate the presence or absence of preconditioning.

The features used as input to the logistic regression model were one-hot encoded. One-hot encoding is a method used to convert categorical variables into a binary matrix representation. Each category is represented by a unique vector where only one element is "1" (active), and all others are "0." This transformation ensures that categorical features can be used in models that require numerical input while avoiding unintended ordinal relationships.

Model training and hyperparameter optimization

Classification Models:

1. Model A: Preclinical data training and validation

- Data Preparation: The preclinical dataset was used to train the model.

Model Definition: A logistic regression model with objective function:

$$J(w) = -\frac{1}{N} \sum_{i=1}^N [y^{(i)} \log(h_x(w^T x^{(i)})) + (1 - y^{(i)}) \log(1 - h_x(w^T x^{(i)}))] + CR(w)$$

Where:

- N is the number of training examples.
- M is the number of training parameters.

- 121 - $y^{(i)}$ is the true label (0 or 1) for the i -th training example.
- 122 - $x^{(i)}$ is the feature vector for the i -th training example.
- 123 - w is the weight vector (parameters).
- 124 - $h_x(w^T x^{(i)}) = 1 / (1 + \exp(w^T x^{(i)}))$ is the logistic (sigmoid) function.
- 125 - C is the regularization strength, which controls the trade-off between
- 126 fitting the data and penalizing large model parameters to prevent
- 127 overfitting. A higher value of C increases the penalty on large
- 128 weights, making the model simpler and less prone to overfitting.
- 129 - $R(w)$ is the regularization function (either $R(w) = L_1 = \sum_{j=1}^M |w_j|$ or
- 130 $R(w) = L_2 = \frac{1}{2} \sum_{j=1}^M w_j^2$). Regularisation penalizes large coefficients,
- 131 preventing the model from fitting noise.

132

133 The following were defined as hyperparameters to tune:

- 134 ○ C : Regularization strength, 100 samples ranging from 1e-4 to 1e4
- 135 ○ *penalty*: ['l1', 'l2']
- 136 ○ *solver*: ['lbfgs', 'newton-cg', 'newton-cholesky']

137 Class weights were balanced to account for label imbalance.

- 138 ● Hyperparameter Optimization: A grid search over the defined parameter
- 139 space with 5-fold cross-validation (CV) was performed (via repeated
- 140 stratified CV, for a total of 50 folds). Grid search is a hyperparameter
- 141 optimization technique that systematically evaluates a predefined set of
- 142 hyperparameter combinations. Each combination is tested using cross-
- 143 validation, and the one yielding the best performance according to a
- 144 specified metric is selected.
- 145 Multiple scoring metrics were used, including Area Under the Curve (AUC),
- 146 Macro F1, accuracy, and average precision. The final model was selected
- 147 based on the best average precision score.
- 148 ● Model Training and Validation: The model was trained and validated using
- 149 a 5-fold cross-validation strategy on the preclinical dataset. This approach
- 150 ensured that the model's performance was assessed on multiple subsets of
- 151 the data, mitigates overfitting and provides a more robust evaluation of its
- 152 generalizability.

153

154 2. Model B: Preclinical data training and clinical data testing

- 155 • Data Preparation: The model was trained on the preclinical dataset and
156 tested on the clinical dataset. Only features common to both datasets
157 ("Solid or Hematologic tumors", 'scFv source', 'Target', 'CAR generation',
158 'TM domain', 'Preconditioning', 'Costimulatory domain') were used to ensure
159 compatibility. The same preprocessing steps as in Model A were applied.
- 160 • Model Definition and Hyperparameter Optimization: The logistic regression
161 model was defined, and hyperparameters optimized as described for Model
162 A. The optimal hyperparameters identified were used to train the model on
163 the preclinical data.
- 164 • Model Testing: The trained model was then tested on the clinical dataset to
165 evaluate its predictive performance on external data. Performance metrics
166 such as the Area Under the Receiver Operating Characteristic Curve (ROC
167 AUC), Macro F1, accuracy, and average precision were calculated.

168

169 3. Model C: Clinical data training and validation

- 170 • Data Preparation: The clinical dataset was used for both training and
171 validation. The same preprocessing steps as in Model A were applied to
172 this dataset.
- 173 • Model Definition and Hyperparameter Optimization: The logistic regression
174 model was defined, and hyperparameters optimized as described for *Model*
175 *A*. The optimal hyperparameters identified were used to train the model on
176 the clinical data.
- 177 • Model Training and Validation: Similar to Model A, a 5-fold cross-validation
178 strategy was employed to train and validate the model on the clinical
179 dataset.

180

181 4. Random model

- 182 • A random baseline model, assigning response labels ("Non-responders"
183 and "Responders") with a 50% probability for each, was implemented for
184 each data subset. These subsets included preclinical data categories: "All
185 tumors," "Hematologic," and "Solid"; and clinical data categories: "All
186 tumors" and "Hematologic." This model served as a benchmark for
187 evaluating classification metrics.

188

189 For each model, performance metrics AUC, macro F1 score, sensitivity, and recall
190 were calculated. The AUC reflects the model's ability to distinguish between
191 positive ("responders") and negative ("non-responders") classes, with higher
192 values indicating better performance. The Macro F1 score, a class-agnostic
193 metric, calculates the harmonic mean of precision and recall, making it particularly
194 useful for imbalanced class distributions. Higher macro F1 values indicate superior
195 performance, demonstrating the models' effectiveness in handling class
196 imbalances in treatment responses. Sensitivity measures the proportion of actual
197 responders correctly identified by the model, and specificity measures the
198 proportion of true non-responders accurately classified.

199 Feature importance was determined by extracting coefficients from a trained
200 logistic regression model. The coefficients for each feature were computed across
201 all response classes, and the absolute values of these coefficients were
202 aggregated by the original feature names. This process allowed for the
203 identification of the most influential features contributing to the model's predictions.

204

205 Subset Analysis

206 Both preclinical and clinical datasets were divided into hematological and solid
207 tumors:

- 208 • Preclinical: 171 solid tumor data points, 132 hematological tumor data
209 points
- 210 • Clinical: 86 hematological tumor data points, 19 solid tumor data points

211 For each subset, models A, B, and C were trained, and the best hyperparameters
212 were identified. The performance metrics and feature importance rankings were
213 computed and reported for the best-performing models.

214

215 Regression Model

216 Model D, a L1-regularized linear regression model (Lasso regression) was trained
217 on the clinical dataset to predict the continuous ORR. The training pipeline
218 consisted of the following steps:

- 219 • Preprocessing: Categorical variables were encoded using one-hot
220 encoding.
- 221 • Model Definition: A linear regression model with objective function:

222

$$J(w) = -\frac{1}{N} \sum (y^{(i)} - w^T x^{(i)})^2 + \lambda \sum_{j=1}^M |w_j|$$

223

Where:

224

- $J(w)$ is the objective function.

225

- N is the number of training examples.

226

- M is the number of training parameters.

227

- $y^{(i)}$ is the true value for the i -th training example.

228

- $x^{(i)}$ is the feature vector for the i -th training example.

229

- w is the weight vector.

230

- λ is the regularization parameter that controls the penalty's strength.

231

232

The following were defined as hyperparameters to tune:

233

○ λ : Regularization strength, 100 samples ranging from 1e-4 to 1e4

234

- Model Training: A Lasso regression model was trained, with hyperparameter optimization performed using a grid search to identify the optimal alpha parameter. The best model was selected based on cross-validation performance, with significant features identified by their non-zero coefficients.
- Evaluation: The model's performance was assessed using the coefficient of determination (R^2), calculated from the relationship between the true and predicted ORR values for each cross-validation fold.
- Feature Analysis: Non-zero feature weights from the Lasso regression were extracted and ranked by their absolute magnitude to determine their relative importance in predicting the ORR.

245

246

Environment

247

The analysis was conducted using a conda environment with Python 3.10.13,

248

scikit-learn 1.3.2, and seaborn 0.13.0 for plotting.

249

The code used for the machine learning analysis is available at

250

<https://github.com/DanScarc/CAR-T-Meta-Analysis.git>.