Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

LLM-assisted large-scale curation of peptide clinical trials maps trends in therapeutic development and trial outcomes

Emily Zhang

ehzhang@bccrc.ca

BC Cancer Research Institute

Uluc Birol

Amphoraxe Life Sciences Inc.

Maya Alev

BC Cancer Research Institute

Iris Caglayan

BC Cancer Research Institute

Emre Demirsoy

BC Cancer Research Institute

Mercan Deniz

BC Cancer Research Institute

Ali Salehi

BC Cancer Research Institute https://orcid.org/0009-0006-9991-724X

Berke Ucar

BC Cancer Research Institute

Anat Yanai

Amphoraxe Life Sciences Inc.

Inanc Birol

The University of British Columbia

Article

Keywords:

Posted Date: August 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10618534/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Yes there is potential Competing Interest. The authors declare the following financial interests, which may be considered as potential competing interests: IB is a co-founder, board member, and executive of Amphoraxe Life Sciences Inc.; UBB is a research and development technician at Amphoraxe Life Sciences Inc; AY is an industrial research scientist at Amphoraxe Life Sciences Inc.; AS, BU, ED, and EZ were funded for industrial internships at Amphoraxe Life Sciences Inc. (IT44733, IT43363, IT44733, and IT43363, respectively). Other authors declare no competing interests that could have influenced the work reported.

Abstract

Interest in peptide-based therapeutics has rapidly grown in recent decades, with peptide drugs promising high target specificity and low immunogenicity. However, structural concerns, such as *in vivo* instability and poor membrane permeability can bottleneck clinical translation, and large-scale insights into peptide drug trial outcomes and design features remain limited. Here, we present and validate an automated, large language model-assisted data curation workflow that demonstrates annotation agreement (AC1 scores) across clinical trial records comparable to volunteer-based annotations, including delivery mode (0.75), peptide sequence (0.70), trial outcome (0.60), and reason-for-failure (0.90) labels. Using this framework, we compile and analyse a comprehensive historical dataset of 6,834 peptide clinical trials. We show that injectable delivery modalities predominate among peptide therapeutics (73.47%), that peptide pharmaceutical trials have phase-wise success rates of 26.36% for Phase 1 trials, 33.40% for Phase 2, and 57.74% for Phase 3, and that peptide drugs fail due to toxicity or safety concerns far less often than a general drug benchmarking set in late-stage trials (4.38% vs. 17.15% for Phase 3 trials). These findings clarify key patterns and challenges in peptide translation and drug development and offer a scalable methodology for large-scale biomedical data curation.

Introduction

Peptide-based therapeutics represent a promising and quickly growing sector of the pharmaceutical industry. Since the introduction of insulin to the pharmaceutical market in 1923, commercial interest in peptide drugs has steadily risen. They currently account for 5–6% of the global pharmaceutical market[1] and are projected to achieve a compound annual growth rate of 10.8% by 2028.[2] Peptides are very versatile drugs, possessing a wide variety of applications, including but not limited to oncology,[2] diagnostics,[2] metabolic disease,[3] and antimicrobial activity.[2] Compared to small molecule drugs and biologics, peptides combine high affinity and specificity with less immunogenicity,[4] driving increased interest into their development.

However, the development of a new therapeutic is a costly and time-consuming endeavour; it requires an estimated mean of $172.7 million USD (2018 dollars)[5] and 7.3 years of active research from the initiation of the first in-human study to marketing authorization.[6] Clinical trials of peptide drugs are similarly prolonged, with one study reporting a mean development timeline of 8.02 years.[7] Despite these substantial investments, success rates for clinical trials, regardless of drug type, are relatively low, with one study projecting that only 13.8% of candidates successfully advance from Phase 1 through regulatory approval.[8]

Key translational challenges in the development of new peptide drugs include their relative instability *in vivo* and poor membrane permeability.[2] However, design features, such as the peptide length and charge, amino acid composition, structural modifications, drug conjugates, and the delivery mode of the drug, can improve their effectiveness and bioavailability.[4] Beyond biological hurdles, clinical trials also

frequently fail or undergo early termination due to insufficient enrolment, strategic business decisions, or a combination of multiple factors. As such, navigating this multi-parametric landscape can be highly complex.

Despite rapid growth in this sector, large-scale analysis of clinical trial outcomes specifically targeting peptide therapeutics remains limited. Critical information, such as peptide sequences, delivery modalities, specific trial outcomes, and reasons for failure, are often unavailable in structured formats within public repositories. While databases like the Clinicaltrials.gov website (https://clinicaltrials.gov), maintained by the National Institutes of Health (NIH), provide a centralized hub of 589,453 clinical trials (as of June 12, 2026), the majority of peptide-related entries lack specific outcome and sequence information. This gap in the data restricts broader inferences regarding the translational challenges associated with peptide pharmaceutical development.

The manual inspection of trial records, publications, company reports, and more document types, requires substantial expert effort, limiting the scope of a large-scale study. These annotation procedures require strict standard-operating-procedures (SOPs) with clear definitions, and annotators with strong contextual knowledge in the field to ensure high-quality annotations, further bottlenecking the process. Even with contextual knowledge, for each trial and annotation field, annotators must systematically examine multiple sources to ensure data completeness and accuracy.

Recent advances in large language models (LLMs) have created new opportunities for large-scale data curation, providing an efficient, automated, and standardized workflow. Previous studies, such as one published by Chen et al. in 2025, describe the use of LLMs to gather a clinical trial outcome dataset,[9] but provide little characterization of annotation quality. In this study, we develop and validate an LLM-assisted workflow for the large-scale annotation of peptide clinical trials, quantitatively comparing LLM-generated and human-curated classifications for delivery mode, peptide sequence, trial outcome, and reason-for-failure labels. We show that LLM-human agreement is not significantly different from human-human agreement. Using this validated workflow, we curate a dataset of 6,834 peptide-based interventional trials from ClinicalTrials.gov spanning more than three decades and characterize trends in peptide therapeutic development, including trial duration, participant enrolment, delivery strategies, peptide physicochemical properties, phase-specific success rates, and primary drivers of trial failure. Together, these analyses position LLM-assisted curation as a scalable alternative to purely manual annotation and provide a data-driven overview of historical and ongoing trends in peptide pharmaceutical development.

Results

Using the Claude Opus 4.8 v2.1.177 model (see Methods), we annotated a dataset of 6,834 clinical trials evaluating peptide-based therapeutics. This dataset includes Phase 1, 2, 3, and 4 trials. There are 1,956 finished trials with known outcomes.

Validation of LLM-assisted annotations against human raters

To establish the reliability of the LLM-assisted curation workflow, we first compared LLM-generated annotations with a human-curated reference set of 610 peptide clinical trials.

Using Gwet's AC1 score,[10] we assessed agreement for delivery mode, peptide sequence, trial outcome, and reason-for-failure labels between Human-Human and Human-LLM rater pairs. The statistical difference in computed AC1 scores between Human-Human and Human-LLM annotation comparisons was assessed for each feature within each split, with p-values adjusted for the 12 comparisons using Benjamini-Hochberg correction.

On the holdout test set, Claude achieved human-LLM agreement comparable to Human-Human agreement across all annotation features (Delivery Mode: AC1 = 0.75 vs. 0.70; Sequence: 0.70 vs. 0.46; Outcome: 0.60 vs. 0.61; Reason for Failure: 0.90 vs. 0.86) (Supplementary Fig. 1). After Benjamini–Hochberg correction, none of the 12 comparisons showed statistically significant difference between Human-Human annotations and Human-LLM ($q < 0.05$; Supplementary Fig. 1).

This indicates that the LLM-assisted workflow can match the consistency of volunteer annotators under our annotation protocol. We therefore treat the LLM-assisted annotations as sufficiently reliable for downstream analyses of development timelines, enrolment patterns, and clinical trial outcomes.

Clinical Trial Design and Development Characteristics

Phase 1 peptide trials required a median of 1.75 (IQR 0.58–4.00) years to complete, Phase 2, 3.00 (IQR 1.61–5.58) years, Phase 3, 2.35 (IQR 1.40–4.86) years, and Phase 4, 2.09 (IQR 1.25–3.41) years (Fig. 1). Across all phases, the overall median duration was 2.33 (IQR 1.17–4.42) years (Fig. 1). Notably, the IQRs are quite large for each of these phases, indicating substantial variability in trial timelines (Fig. 1).

Several outliers were observed. The longest trial spanned 32.69 years (Fig. 1); this Phase 1 trial ran from 1997 to an estimated completion date of 2029, investigating how various mechanistically unique medications control excessive increases in heart rate with standing.[11] Conversely, the shortest trial completed in just 18 days in 2020.[12] Phase 4 clinical trials had the lowest upper bound in trial duration of 23.95 years (Fig. 1). Phase 4 (post-approval/post-licensure and post-marketing studies/surveillance studies)[13] also has the narrowest IQR of 1.25 to 3.41.

Participant enrolment rates varied substantially across both clinical phases and trial outcomes (Fig. 2). Clinical trials with positive outcomes were associated with significantly higher enrolment rates, with a median of 220.5 participants, compared to 43.0 for negative outcome trials (Mann–Whitney $U = 534{,}574$, $P < 0.001$) (Fig. 2). This trend was relatively consistent across all clinical phases despite temporal variation, indicating that successful trials are often associated with larger participant cohorts.

The largest study identified was a Phase 3 trial with 13,299 participants (Fig. 2), which evaluated the efficacy and safety of Tirzepatide on patients with Type 2 Diabetes that were experiencing major adverse

cardiac events.[14] Each phase showed substantially different median enrolments, with Phase 1 median being 19.00, Phase 2 being 47.00, Phase 3 being 345.00, and Phase 4 being 72.00 (Fig. 2). This trend is consistent with the objectives of late-stage clinical development, where larger populations are often required to evaluate efficacy and safety in more diverse groups.[13]

Peptide Therapeutic Design

Of the drugs being tested in the clinical trials, 73.47% are delivered by injection or infusion (Table 1). This includes but is not limited to: intravenous infusion, subcutaneous-, intradermal-, intramuscular-, and intravitreal injections. Oral delivery represented the second most common route for peptide-based drugs, accounting for 14.82% of the trials (Table 1). This includes delivery via tablets, foods, and drinks. Topically-delivered drugs, including via cream, powder, covering, sprays, and washes, are the least common broader category, making up 4.77% of trials (Table 1). Some alternative delivery modalities include inhalation and intranasal routes.

Table 1
Delivery modes of the peptide-based drugs being tested in the clinical trial dataset.

Delivery Mode	Count
Injection/Infusion	5,021 (73.47%)
Oral	1,013 (14.82%)
Topical	326 (4.77%)
Other	474 (6.94%)

Across all trial outcomes, single-sequence peptide drugs were more common than multi-sequence drugs (Fig. 3A). In active trials, there were 3.83X more trials testing single-sequence drugs than multi-sequence (Fig. 3A). The single-sequence to multi-sequence ratio was similarly high for terminated trials (2.51-fold) and withdrawn trials (2.89-fold), but dropped closer to parity in positive-outcome and failed trials (1.36-fold and 1.84-fold, respectively) (Fig. 3A).

Positive-outcome trials showed a relatively narrower distribution in peptide length, with an IQR of 26.50 to 32.00. Otherwise, the distribution of sequence lengths of the evaluated peptides was relatively similar across trial outcomes, ranging from 20.37 amino acids at the shortest mean length for terminated trials to 29.60 at the longest for successful trials (Fig. 3B). For multi-peptide drugs, the mean length and charge for all of the sequences in the cocktail was taken before computing overall means across the outcome categories. In contrast to length, the mean peptide charge differed substantially between the different clinical outcomes; terminated trials had a mean charge of 0.81 and failed trials had a mean charge of 0.44 (Fig. 3C). The mean charge of positive trials (0.50) was relatively close to successful trials (Fig. 3C).

Lastly, we observed an almost 1:1 ratio of natural to unnatural peptides in active trials (0.84:1), which differs substantially from the ratio of 2.28:1 for natural to unnatural peptides for successful trials and 3.59:1 for withdrawn trials (Fig. 3D). We classified a peptide sequence as unnatural based on the presence of d-amino acids, N- and C-terminus modifications, cyclization, PEGylation, and conjugation to other drug types.

Taken together, these observations suggest candidate features that may be associated with clinical trial outcome and could inform future predictive modeling efforts for peptide trial success.

Clinical Trial Outcomes and Reasons for Failure

For Phase 1 trials, 26.36% of trials demonstrated positive outcomes, 27.27% were withdrawn, 38.64% were terminated, and 7.73% failed (Fig. 4A). Success rates increased in Phase 2, with 33.40% successful trials, 15.51% withdrawals, 34.39% terminations, and 16.70% failures (Fig. 4A). Phase 3 trials showed the highest success rate, with 57.74% of trials reporting positive outcomes (Fig. 4A), alongside the lowest withdrawal rate, at 4.39% (Fig. 4A). Lastly, Phase 4 trials had a success rate of 46.08%, with withdrawal, termination, and failure rates standing at 10.06%, 26.96%, and 16.90%, respectively (Fig. 4A).

We also compared the outcome of peptide-specific clinical trials against a multi-drug type clinical trial cohort published by Wong et al. (2019), which detailed phase-by-phase probabilities of success for 185,994 clinical trials between January 1, 2005 and October 31, 2015.[8] Within this overlapping time period, the dataset that we curated captures 806 peptide-specific, pre-approval trials, with 118 Phase 1 trials, 269 Phase 2 trials, and 419 Phase 3 trials. The overall success rates are the product of the success rates across the three consecutive phases for each calendar year. Peptide clinical trials showed success rates that were generally comparable with the multi-drug background dataset across all three phases (Fig. 4B).

The dataset contained 1,211 trials that were terminated, withdrawn, or completed with reported failed results (Table 2). Clinical inefficacy was the largest reason for failure, accounting for 34.68% of all failed trials (Table 2). The next largest reason for failure was commercial or business constraints, which can include funding concerns, strategic company decisions, and shifts in research focus for principal investigators; these accounted for 27.09% of the failed trials. Recruitment issues were the next most common reason (25.52%), followed by safety concerns (4.38%) and disruptions from the COVID-19 pandemic (1.82%) (Table 2). These associations do not by themselves establish causality and are likely influenced by unmeasured design and context variables.

Table 2
Primary reasons for peptide clinical trial termination, withdrawal, and failure.

Reason for Failure	Phase 1	Phase 2	Phase 3	Phase 4	Unknown Phase Label	Overall
Ineffective for intended purpose	20	100	153	93	54	420
Business reasons	67	92	65	65	39	328
Recruitment issues	46	93	60	84	26	309
Toxicity	4	23	16	8	2	53
Due to COVID-19 pandemic	5	9	4	3	1	22
Unknown	20	18	13	15	13	79

While inefficacy was the most common reason for failure for Phase 2–4 trials, Phase 1 trials are primarily designed to assess safety and feasibility rather than efficacy; accordingly, inefficacy accounted for only 12.35% of Phase 1 failures, whereas commercial reasons and recruitment-related concerns were more common (Table 2).

Discussion

In this study, we developed and validated an LLM-assisted workflow for the large-scale annotation of peptide clinical trials and applied it to curate a dataset comprising 6,834 peptide studies spanning over three decades. This enabled the characterization of historical trends in peptide therapeutic development, including clinical development timelines, participant enrolment patterns, delivery strategies, peptide sequence characteristics, trials success rates, and reasons for trial failure. Our analysis demonstrates that LLMs can achieve human-level data extraction fidelity when evaluating multiple data sources, highlighting its potential as a scalable approach for biomedical data curation.

Our results show that peptide clinical trials require, on average, 1.75 (IQR 0.58–4.00) years to complete Phase 1, 3.00 (IQR 1.61–5.58) years for Phase 2, and 2.35 (IQR 1.40–4.86) years for Phase 3 (Fig. 1). Thus, active trial phases collectively require a median of approximately 7.10 years (Fig. 1), excluding time between phases, regulatory approval process, and preclinical research. This timeline is broadly consistent with established literature, which estimates that the average time in the United States for a drug to move from pre-clinical testing to Food and Drug Administration approval is 12 years.[15] This finding is consistent with a previously published study that isolated different drug modalities and reported an average development time of 8.02 years for peptide drugs.[7]

Another notable insight from our dataset is the continued dominance of injectable delivery strategies; more than 73% of all recorded trials used injection or infusion-based administration routes (including intravenous infusion, subcutaneous injection, intradermal injection, intramuscular injection, and

intravitreal injections) (Table 1). This trend likely reflects longstanding challenges associated with peptide therapeutics, including instability *in vivo* and limited bioavailability.[2] As such, the most common drug delivery mode identified in this dataset is injection or infusion (Table 1), as this modality can circumvent some of these barriers.[16] However, this can be associated with poor patient compliance due to discomfort, dislike of needles, and irritation.[17]

In response to the challenges associated with injection-based administration, there has been a substantive research focus on the optimization of oral delivery technologies and other non-invasive formulations in recent years.[18] The relatively recent approval of oral semaglutide marks a substantial milestone in this field of research.[18] In our curated dataset, 14.82% of the drugs are delivered orally (Table 1), indicating meaningful investment in this field. Despite this progress, additional research is required to overcome the remaining challenges associated with the oral delivery of peptide drugs.[18]

Drug delivery modality was significantly associated with trial success rate ($\chi^2 = 21.47$, n = 2,189, P = 0.01), with injected and orally administered drugs demonstrating higher success rates than topically delivered drugs and other delivery modalities (Supplementary Fig. 2). However, no significant association was observed between delivery modality and the proportion of trials failing due to either lack of efficacy or safety concerns when analysis was restricted to these failure categories. Together, these findings suggest that although delivery modality may influence overall clinical trial success, it is not strongly linked to specific mechanisms of trial failure (Supplementary Fig. 2).

The physicochemical properties of peptide therapeutics represent another important consideration in drug development. Peptides typically exhibit high hydrophilicity and hydrogen bonding capacity, limiting passive membrane transport and their bioavailability.[2] The sequence length and charge of peptides have been indicated as determinants of peptide stability and permeability.[19] For example, increasing peptide length can sharply decrease membrane permeability, which could negatively impact the efficacy of the drug.[20] However, our dataset demonstrates relatively consistent mean sequence lengths across all trial outcomes, including successful and failed candidates (Fig. 3B). Charge distributions, on the other hand, demonstrated substantial overlap and variability, albeit with no clear conclusion arising from the data (Fig. 3C). For example, the mean charges for failed trials and successful trials (0.44 and 0.50, respectively) are closer to each other than they are to the mean charge for terminated trials (0.81) (Fig. 3C). As such, while there is established literature on the impact of peptide length and charge on the stability of therapeutics *in vivo*, it is difficult to draw any conclusions with this dataset. These simple descriptors do not capture higher-order biological features, such as secondary structure, target binding affinity, stability, immunogenicity, and more. Consequently, predictive models may require higher-order structural and biochemical features.

To address the limitations of natural peptide sequences, several engineering strategies have been developed to enhance stability, such as the substitution of existing amino acids with other amino acids or D-stereoisomers, acylation, PEGylation, and cyclization.[16] Such modifications can improve resistance

to proteolytic degradation, prolong drug half-life, enhance tissue distribution, and reduce immunogenicity.[21–25] However, despite these established advantages, only 36.36% of the trials in our dataset evaluated peptide drugs with non-canonical amino acids or sequence modifications, and only 30.48% of trials with positive outcomes tested peptides with unnatural amino acids (Fig. 3D). Broader adoption of these sequence modifications could improve on the historically poor bioavailability and half-life of peptide therapeutics, improving clinical success rates across all delivery modalities, but particularly non-invasive delivery routes.

Beyond drug design, clinical trial design and cohort scale can be factors associated with clinical outcomes. Across all development phases, trials with positive outcomes, i.e. demonstrated efficacy and/or safety *in vivo*, were often associated with higher median participant enrolment (Fig. 2). However, this association is likely multifactorial: trials with higher cohorts may indicate more funding, more resources, and stronger trial infrastructure, leading to more successful outcomes. Conversely, failed trials might terminate early before enrolling all participants, decreasing their total enrolment counts. In our analysis, we removed trials with zero participants but left in all other trials. Between the phases, we generally see higher mean enrolment rates for Phase 3 and 4 trials compared to Phase 1 and 2, which is expected, given the different scopes for each trial phase.[13] The phases do not show any notable temporal trends in enrolment for trials with positive nor negative outcomes (Fig. 2).

In examining trial success rates, our peptide-specific dataset showed phase-by-phase and cumulative success rates that were comparable to the multi-drug benchmark reported by Wong et al. (2019)[8] for studies initiated between 2005 and 2015 (Fig. 4B). In our dataset, spanning 1987 to 2026, the Phase 1 success rate was 26.36%, and 33.40% and 57.74% for Phases 2 and 3, respectively, corresponding to an estimated probability of eventual approval of 5.08% given phase-wise success (Fig. 4A). The highest failure rate was observed in Phase 3 clinical trials (19.43%; Fig. 4A). This may reflect the larger scale and greater complexity of late-stage studies, which require substantial participant enrolment, funding, and operational resources.[26] Several studies that appear promising in early phases fail to demonstrate benefit or safety in larger, late-stage settings,[26] and the larger, more heterogeneous participant cohorts may also increase the likelihood of detecting unforeseen adverse events. This is particularly important in Phase 3, where studies should be adequately powered to detect even common adverse events.[27] The higher failure rate in Phase 3 is therefore notable, not only due to the cost of the study itself, but also because of the time and financial investment already committed in earlier phases.[28]

The analysis of this dataset can also provide additional insights into the common reasons for failure for peptide clinical trials. We have identified that the majority of peptide trials fail due to inefficacy for its intended purpose and that relatively few fail due to safety concerns (Table 2). This contrasts with previously published literature. For instance, one study published by Hwang et al. in 2016 reported that in a sample of 344 failed Phase 3 trials between 1998 and 2015, testing various drug types, 17.15% of them failed due to safety concerns.[26] This differs substantially from our dataset, which finds that of 311 Phase 3 peptide trials with failed outcomes, only 5.14% failed due to toxicity (Table 2). Furthermore, that

same study found that 56.69% of Phase 3 trials failed due to lack of efficacy,[26] compared with 49.20% in our dataset (Table 2). Peptide clinical trials do share similar rates of failure due to commercial reasons compared with the general benchmarking dataset (20.90% and 21.50%, respectively).[26]

Another major contribution of this study is the evaluation of the use of Claude and locally-installed LLMs for automated large-scale curation of clinical trial data. Each annotation strategy offers distinct advantages and limitations. Claude benefits from access to a broad range of external information sources, which may improve its ability to resolve ambiguous or incomplete trial records. In contrast, locally installed LLMs are restricted to predefined source materials provided during the data collection stage. Although this limitation may reduce access to potentially relevant information, it also provides greater control over source quality and may reduce the influence of unreliable or inconsistent references. Claude additionally offers advantages in computational efficiency and annotation speed, whereas locally installed LLMs require substantially greater processing time, computational resources, and storage capacity (Supplementary Table 6).

Conversely, locally deployed LLMs may offer superior reproducibility. Commercial LLMs such as Claude, are frequently updated, and changes in model architecture, training data, or system behavior may introduce variation in annotation outputs over time. By comparison, locally installed models can be version-locked and executed within a controlled computational environment, enabling more consistent and reproducible annotations across longitudinal curation efforts. These findings suggest that the optimal annotation strategy depends on the relative importance of efficiency, source flexibility, and reproducibility within a given research context.

In summary, we present a scalable methodology to annotate peptide clinical trials using LLMs, providing a less time-intensive process for curating a dataset. Notably, this methodology enabled us to find the outcome of peptide clinical trials and the sequences of the therapeutics, allowing us to provide novel insights within this field. This methodology produces interrater agreement scores that are not significantly different from Human-Human interrater agreement, suggesting its reliability as an annotation methodology relative to the *status quo*. With this methodology, we curated dataset of 6,834 peptide clinical trials, allowing us to examine current and historical trends in the pharmaceutical development space.

Methods

ClinicalTrials.gov Data

The unannotated peptide clinical trial records were directly downloaded from the ClinicalTrials.gov website. The search term ‘peptide’ was used in the ‘Other Terms’ window and the ‘Study Type’ was filtered to only include ‘Interventional’ studies. This search was performed on June 12, 2026 and returned 37,648 trials. The specific fields that were downloaded are summarized in Supplementary Table 1.

Manual Annotation Workflow

To establish a baseline dataset for validation, a subset of 864 clinical trials were systematically annotated by human volunteers. The annotators were informed of all the fields summarized in Supplementary Table 1 prior to annotation. The annotators were instructed to determine: (1) whether or not the interventional drug is a peptide, since not all of the returned trials were testing peptide-based interventional drugs; (2) the delivery method of the drug; (3) the complete amino acid sequence(s) of the peptide therapeutics; (4) the outcome of the clinical trial; and (5) the reason for failure if the trial was classified as 'Failed', 'Withdrawn', or 'Terminated'. The categorical classifications of the annotated features are summarized Supplementary Table 2.

The annotators were given a standard operating procedure (SOP) detailing the annotation workflow prior to beginning (Supplementary Text 1). Annotators were also required to provide explicit URL links to evidence supporting each annotation entry.

To quantify human baseline variability, each clinical trial was annotated twice in a blinded fashion – once by the project lead (designated as Replicate 1) and once by a volunteer (designated as Replicate 2). The trials in replicate 2 are non-overlapping (i.e. no two volunteers annotated the same trials). The volunteers each annotated different numbers of trials based on their individual capacity; the division of labour is summarized in Supplementary Table 3. The volunteers also represent a range of educational backgrounds (Supplementary Table 3).

Of the human-annotated clinical trials, 610 were annotated as testing peptide-based drugs.

LLM-based Annotation Workflow

LLM-based annotations were generated using the Claude Opus 4.8 v2.1.177 model (Anthropic, Max effort setting).[32] In the initial extraction phase, Claude was provided with a prompt to parse 37,648 trials downloaded from ClinicalTrials.gov, selecting only valid interventional peptide drug trials (Supplementary Table 5). This step yielded 6,834 unique records and their corresponding National Clinical Trial (NCT) identifiers were saved in a CSV file. The Claude-based annotation process was conducted on June 13th and June 14th, 2026.

The target annotations fields (Supplementary Table 2) were annotated in separate sub-directories and chats, following the chain-of-thought prompting methodology, to allow for more complex reasoning.[33] The prompts to generate the initial annotations are detailed in Supplementary Table 5. Following initial annotation generation, the raw results were consolidated into a single file and passed back through the LLM for an additional verification step.[34] We compared the LLM agreement with the human baseline annotations and presented the results in Supplementary Fig. 2.

The LLM-curation workflow is driven by two Python scripts, gather_data.py, to collect information about the clinical trials from a variety of sources, and classify_trials.py, to inform the annotation decision making. Additionally, the LLM is provided with a README file containing broader annotation instructions. Each of the four annotation features contains their own versions of these three files, specifying unique

instructions for each stage of the annotation process. While initial versions of the files were developed by the co-authors of this study, the files were iteratively optimized by the LLM itself during the prompt engineering process. The validation and test datasets were held out during this optimization process. The files used to generate the annotated dataset are available freely on GitHub (https://github.com/BirolLab/LLM-Clinical-Trial-Annotator). Through several iterations, LLM was asked to modify the annotation approach through the modification of the README file and the Python scripts to improve the agreement with the first replicate of human annotations for each trial on the training set. The prompts used during this process can be found in Supplementary Table 5.

An alternative workflow was developed in parallel using a panel of free, locally-installed LLMs, including Llama 3.1:8b,[29] Gemma 3:12b,[30] Qwen3:8b,[31] and Qwen3:14b.[31] This workflow is a three-stage pipeline that includes a research phase, an annotation phase, and a verification phase. Stage 1, the research phase, involves querying the ClinicalTrials.gov Application Programming Interface (API) to extract drug and peptide names associated with the trial ID number. Then, using that extracted information, it queries 20 resources (Supplementary Table 4) to gather information used for trial annotation, saving it in JavaScript Object Notation (JSON) format. Next, in stage 2, that information is parsed and passed to a Qwen3:14b LLM-based annotation agent to annotate the delivery mode and sequence of the peptide drug and the outcome and reason for failure (if applicable) of the trial. The last stage is verification, where three independent models, Gemma 3:12b, Qwen 3:8b, and Llama 3.1:8b, evaluate the correctness of the initial annotation. Gemma 3 is tasked to act more conservatively (defaults to safest interpretation when evidence is ambiguous), Qwen 3 is tasked to only cite explicit facts from sources when validating, and Llama 3.1 is asked to challenge assumptions and edge cases as an adversarial reviewer. At least two of the three annotators must agree with the primary annotation for the annotation to be accepted. If this does not occur, Qwen 3:14b acts as the final reconciler and evaluates all of the evidence provided by the models to make the final decision. The versions of all of the models are described in Supplementary Table 6.

Inter-rater Agreement Analysis

To evaluate the generalizability of the LLM-assisted annotation workflow, the subset of peptide drug clinical trials that were manually annotated by volunteers (610 trials) was separated into stratified train, validation, and test sets with an 80%:10%:10% split with the train_test_split function in Python's scikit-learn package.[35] The stratified split has proportional representations of each volunteer within the train and the validation set. We then conducted inter-rater agreement analysis to assess the quality of LLM annotations.

Gwet's AC1 score[10] was used to quantify inter-rater agreement for the four annotation features (Delivery Mode, Sequence, Outcome, and Reason for Failure). The AC1 score was computed independently for each of the annotated fields for the Human-vs-Human pairs (Replicate 1 vs Replicate 2) and the Human-vs-LLM pairs (Replicate 1 vs LLM) (Supplementary Fig. 3). Trials that only existed in one of the datasets and not the other (human-annotated or LLM-annotated) were removed before computing AC1 scores.

The AC1 score was computed with resampling with 5000 bootstraps to compute 95% confidence intervals. Pre-processing was performed on the 'Sequence' feature for each of the trials before computing the AC1 scores. Null values are standardized and the sequences of multi-sequence drugs are alphanumerically ordered before looking for exact matches. When recording peptide sequences, LLMs and human annotators were sometimes able to capture partial sequence matches. However, when calculating AC1 scores, we only considered the annotations with a full sequence match; most of the sequences were captured by a complete mismatch or full match. The sequence similarity within human annotations and between human and LLM annotations are visualized in Supplementary Fig. 4.

AC1 comparisons were conducted independently for each of the four annotation features across all three data splits (12 comparisons in total), and raw p-values from the bootstrap difference tests were adjusted using the Benjamini-Hochberg procedure to control the false discovery rate at 0.05.[36] This correction was applied across the full family of feature and split comparisons rather than to each split individually, since all 12 tests feed into the same overall conclusion about LLM annotation quality.

For the Claude-based annotations, the statistical difference in computed AC1 scores between Human-Human annotation comparisons and Human-LLM were non-significant in the train, validation, and test sets (Supplementary Fig. 2). This suggests that the LLMs perform relatively as consistently as Replicate 2 (the volunteer annotations), when compared with Replicate 1. However, for the local LLM-based annotations, the Human-LLM Sequence annotation AC1 score was significantly higher than the Human-LLM score at the 0.05 significance level. As a result, the Claude-based annotation workflow was used to generate the final dataset.

Two one-sided tests (TOST) for equivalence were conducted on the bootstrap-resampled AC1 score differences ([Human-Human] – [Human-LLM]) for each annotation feature within each split. The equivalence margin was set at $\Delta = 0.1$, corresponding to half the width of a single agreement category in the Landis and Koch, and Altman and Fleiss interpretation scales (which partition agreement into 0.20-wide intervals). Equivalence was concluded when the 90% bootstrap confidence interval for the AC1 difference fell entirely within $(-\Delta, +\Delta)$. TOST did not establish statistical equivalence for 10 of 12 feature × split comparisons at this margin (Supplementary Fig. 5). Notably, two of the 10 non-equivalent comparisons favoured higher Human-LLM AC1 scores (Supplementary Fig. 5). Given the modest and uneven per-annotator sample sizes underlying several comparisons, these results likely reflect limited statistical power to confirm equivalence within the specified margin rather than evidence of a true difference; AC1 point estimates for Human-LLM agreement remained within 0.1 points of Human-Human agreement for 9 of 12 comparisons.

Data Pre-processing

Some trials were classified as Early Phase 1, Phase 1 and 2, and Phase 2 and 3. Trials that were categorized as Early Phase 1 were grouped with Phase 1 trials, trials that were categorized as multi-phase were grouped with the later phase. Further, any trials that were recruiting participants and active but not recruiting participants during the time of this study analysis were categorized as an active

clinical trial. The sequences of the peptides were not cast to upper or lower case, as lower-case letters were used to denote the d-conformation.

Statistical information

A two-tailed test was used to measure the statistical significance of the difference between AC1 scores associated with the Replicate 1 - Replicate 2 annotations and the Replicate 1 - LLM-based annotations. To assess differences in agreement between human–human and human–LLM pairs, a nonparametric bootstrap procedure (5,000 iterations) was performed on the same subset of observations. In each iteration, AC1 was recalculated for both pairs and the difference ($\Delta AC1 = AC1_{Human-Human} - AC1_{Human-LLM}$) was recorded. The 95% confidence interval for the difference was derived from the 2.5th and 97.5th percentiles of the bootstrap distribution. Two-sided p-values were computed using a bootstrap-based approach by centering the distribution of differences at the observed value and estimating the proportion of resampled differences with magnitude greater than or equal to the observed difference. Statistical significance was categorized using conventional thresholds ($p < 0.05$, $p < 0.01$, $p < 0.001$).

Declarations

Contributions

I.B. conceived the study. E.Z. and U.B.B. implemented the methodology. E.Z analysed the data. E.Z. created figures and tables with input from co-authors. E.Z., and I.B. wrote the manuscript. I.B. supervised the research. M.A., I.C., E.D., M.D., A.S., B.U., A.Y., E.Z. manually annotated the clinical trials. All authors commented on the manuscript.

Competing interests

The authors declare the following financial interests, which may be considered as potential competing interests: IB is a co-founder, board member, and executive of Amphoraxe Life Sciences Inc.; UBB is a research and development technician at Amphoraxe Life Sciences Inc; AY is an industrial research scientist at Amphoraxe Life Sciences Inc.; AS, BU, ED, and EZ were funded for industrial internships at Amphoraxe Life Sciences Inc. (IT44733, IT43363, IT44733, and IT43363, respectively). Other authors declare no competing interests that could have influenced the work reported.

Acknowledgements

This study is supported by Michael Smith Health Research BC-Mitacs Industry-Based Opportunity funding (IT44733) and Mitacs Accelerate Program (IT43363). The funding organizations did not have any role in the design of the study, in the collection, analysis, and interpretation of the data, or in writing the manuscript.

Code Availability

The code used to generate the locally-installed LLM-based clinical trial annotations and the prompts and Python scripts used to generate the Claude-based annotations are available at https://github.com/BirolLab/LLM-Clinical-Trial-Annotator and released under the GPL-3 license.

Data Availability

The annotated clinical trials (by human annotators, the Claude Opus 4.8 v2.1.177 model, and by the panel of locally installed LLMs, Llama 3.1:8b, Gemma 3:12b, Qwen3:8b, and Qwen 3:14b) have been deposited in Zenodo at https://doi.org/10.5281/zenodo.20753318.

References

1. Cresti L, Cappello G, Pini A (2024) Antimicrobial Peptides towards Clinical Application—A Long History to Be Concluded. Int J Mol Sci 25:4870
2. Xiao W et al (2025) Advance in peptide-based drug development: delivery platforms, therapeutics and vaccines. Signal Transduct Target Ther 10:74
3. Teng B, Li J, Ren P (2022) Peptide drugs application in metabolic diseases and discovery strategies. J Holist Integr Pharm 3:24–31
4. Wang L et al (2022) Therapeutic peptides: current applications and future directions. Signal Transduct Target Ther 7:48
5. Sertkaya A, Beleche T, Jessup A, Sommers BD (2024) Costs of Drug Development and Research and Development Intensity in the US, 2000–2018. JAMA Netw Open 7:e2415445
6. Möhrle JJ (2024) How long does it take to develop a new drug? Lancet Reg Health - Eur 43:100998
7. Brown DG, Wobst HJ, Kapoor A, Kenna LA (2022) Southall, N. Clinical development times for innovative drugs. Nat Rev Drug Discov 21:793–794
8. Wong CH, Siah KW, Lo AW (2019) Estimation of clinical trial success rates and related parameters. Biostatistics 20:273–286
9. Chen J et al (2025) TrialBench: Multi-Modal AI-Ready Datasets for Clinical Trial Prediction. Sci Data 12:1564
10. Gwet K, Inter-Rater R (2002) Dependency on Trait Prevalence and Marginal Homogeneity. Stat Methods Inter-Rater Reliab Assess 2
11. US National Library of Medicine (2024) ClinicalTrials.gov https://clinicaltrials.gov/study/NCT00001481
12. *US National Library of Medicine. ClinicalTrials.Gov Https://Clinicaltrials.Gov/Study/NCT05271136 (2020).*

13. Kandi V, Vadakedath S (2023) Clinical Trials and Clinical Research: A Comprehensive Review. Cureus 15:e35077
14. US National Library of Medicine (2023) ClinicalTrials.gov https://clinicaltrials.gov/study/NCT02389036
15. Ogbaudu E, Ohaya TC, Smith JF, Elahi MA Phases of clinical trials. in *Translational Orthopedics* 295–299 (Elsevier, 2024). 10.1016/B978-0-323-85663-8.00072-6
16. Renukuntla J, Vadlapudi AD, Patel A, Boddu SHS, Mitra AK (2013) Approaches for enhancing oral bioavailability of peptides and proteins. Int J Pharm 447:75–93
17. Nicze M et al (2024) The Current and Promising Oral Delivery Methods for Protein- and Peptide-Based Drugs. Int J Mol Sci 25:815
18. Niazi SK (2026) Oral delivery of peptides and proteins: pharmacokinetic boundaries, negative selection, and route triage. Front Drug Deliv 6:1776167
19. Zapadka KL, Becher FJ, Dos Santos G, A. L., Jackson SE (2017) Factors affecting the physical stability (aggregation) of peptide therapeutics. Interface Focus 7:20170030
20. Nielsen AL et al (2025) Bulk Measurement of Membrane Permeability for Random Cyclic Peptides in Living Cells to Guide Drug Development. Angew Chem Int Ed 64:e202500493
21. Al Musaimi O, Lombardi L, Williams DR, Albericio F (2022) Strategies for Improving Peptide Stability and Delivery. Pharmaceuticals 15:1283
22. Gao X, Mazière AD, Beard R, Klumperman J, Hannoush RN (2021) Fatty acylation enhances the cellular internalization and cytosolic distribution of a cystine-knot peptide. iScience 24:103220
23. Trier S, Linderoth L, Bjerregaard S, Andresen TL, Rahbek UL (2014) Acylation of Glucagon-Like Peptide-2: Interaction with Lipid Membranes and In Vitro Intestinal Permeability. PLoS ONE 9:e109939
24. Veronese FM, Mero A (2008) The Impact of PEGylation on Biological Therapies: *BioDrugs* 22, 315–329
25. Vu QN et al (2021) Cyclisation strategies for stabilising peptides with irregular conformations. RSC Med Chem 12:887–901
26. Hwang TJ et al (2016) Failure of Investigational Drugs in Late-Stage Clinical Development and Publication of Trial Results. JAMA Intern Med 176:1826
27. Wahab IA, Pratt NL, Kalisch LM, Roughead EE (2013) The detection of adverse events in randomized clinical trials: can we really say new medicines are safe? Curr Drug Saf 8:104–113
28. Fogel DB (2018) Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review. Contemp Clin Trials Commun 11:156–164
29. Grattafiori A et al (2024) The Llama 3 Herd of Models. Preprint at https://doi.org/10.48550/ARXIV.2407.21783
30. Gemma, Team et al (2025) Gemma 3 Technical Report. Preprint at https://doi.org/10.48550/ARXIV.2503.19786

31. Yang A et al (2025) Qwen3 Technical Report. Preprint at https://doi.org/10.48550/ARXIV.2505.09388
32. Anthropic (2026) Claude 4.6 Opus [Large language model]. https://claude.ai/new
33. Wei J et al (2022) Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Preprint at https://doi.org/10.48550/ARXIV.2201.11903
34. Dhuliawala S et al (2023) Chain-of-Verification Reduces Hallucination in Large Language Models. Preprint at https://doi.org/10.48550/ARXIV.2309.11495
35. Pedregosa F et al (2011) Scikit-learn: Machine Learning in Python. J Mach Learn Res 12:2825–2830
36. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B Stat Methodol 57:289–300

Figures

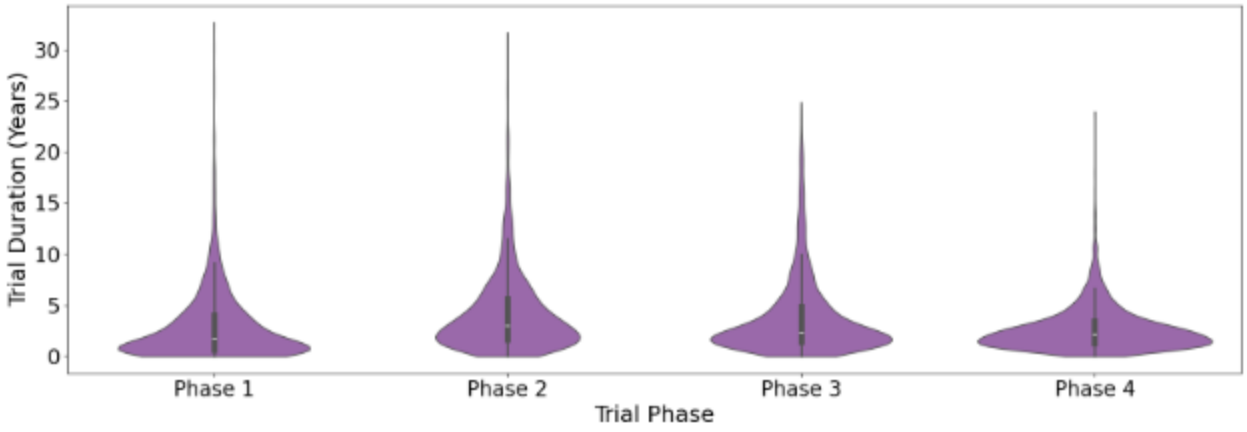


Figure 1

Clinical development timelines for peptide therapeutics across development stages. Violin plots show the distribution of trial duration (years) for peptide therapeutic clinical trials across four development phases (Phase 1–4; n = 5,631). For ongoing trials, the anticipated completion date is used to determine trial duration. Median and interquartile range are indicated by the internal box plot. Trials with missing start or completion dates were removed before data visualization.

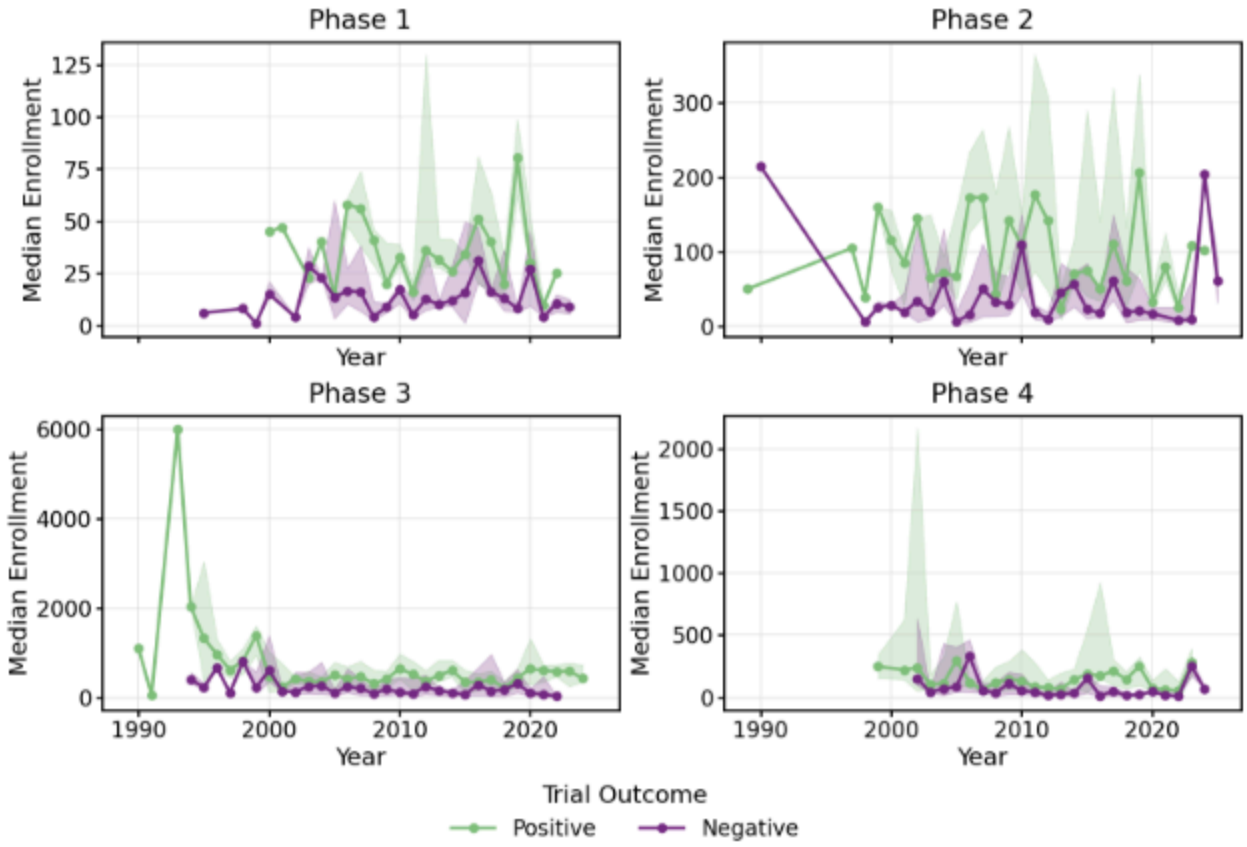


Figure 2

Historical median peptide clinical trial enrolment (number of participants) across development stages. Solid lines indicate the median enrolment, and shaded bands represent the interquartile range (IQR; 25th–75th percentile). The results are stratified by trial outcome with positive outcomes corresponding to trials that met their designated clinical endpoints and safety targets and negative outcomes corresponding to trials that were terminated, withdrawn, or completed but unable to meet their clinical endpoints or demonstrate patient safety. The enrolment is attributed to the start date of the trial. Trials with zero participant enrolment were removed before plotting.

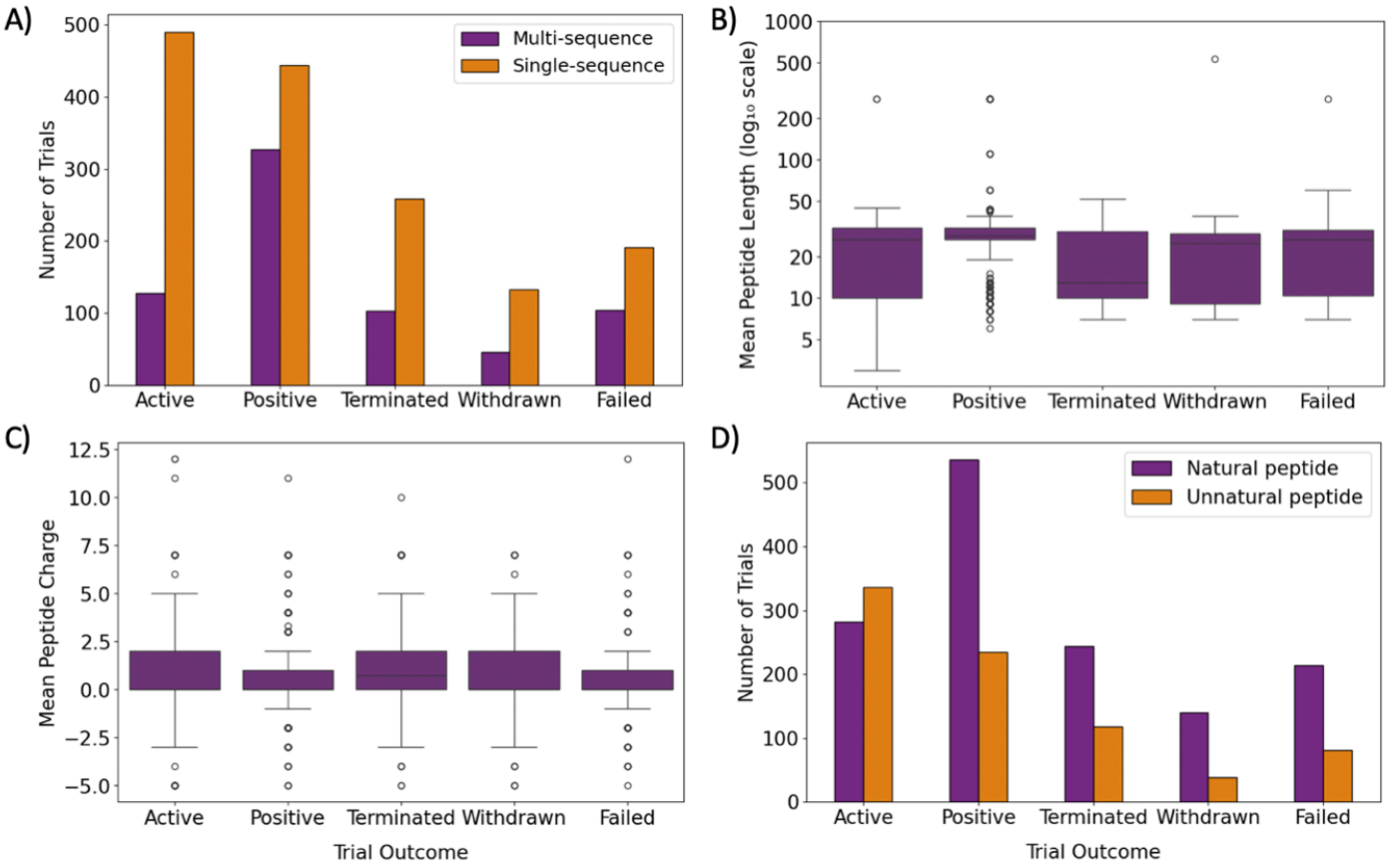


Figure 3

Characterization of peptide sequence complexity, length, net charge, and chemical modifications. A. Comparative classification of single-sequence versus multi-sequence peptide cocktail drugs across different trial outcomes. **B.**Profiling of mean sequence length (log-transformed), **C.** net peptide charge, and **D.**distribution of natural versus unnatural (chemically modified) peptide sequences stratified by trial outcome.

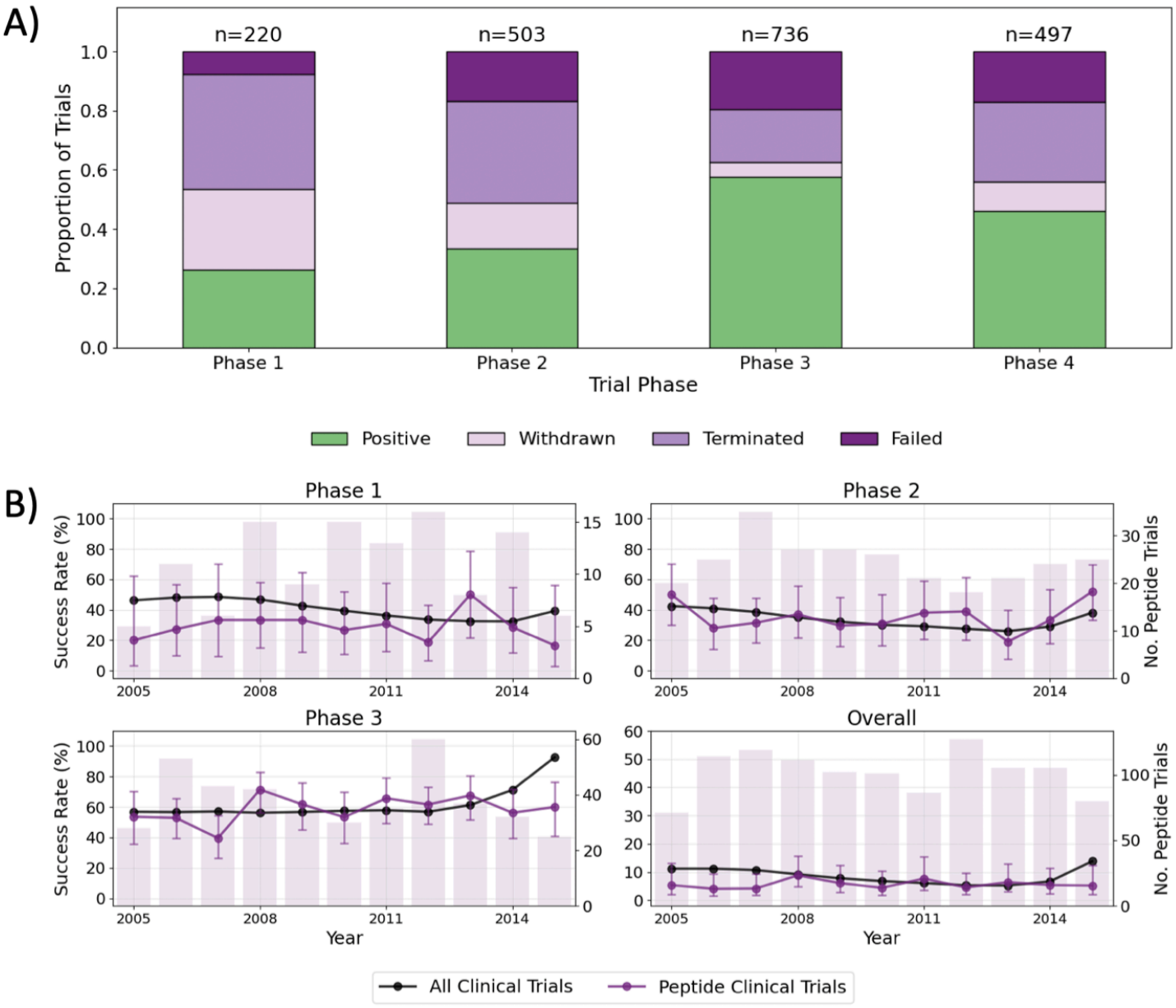


Figure 4

Phase-by-phase clinical trial outcomes and benchmarking of peptide therapeutic success rates. A. Percentage distribution of clinical trial outcomes stratified across Phases 1–4 for the curated peptide cohort (n = 1,956). Trials without a known outcome or phase designation were removed before visualization. **B.** Temporal trends in overall clinical success probability of peptide therapeutics (primary y-axis) compared to an unselected, multi-drug background established by Wong et al. (2019). The number of trials for peptide drugs is represented as a bar chart (secondary y-axis).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ezhangPCTsupplementaryjul27IB.docx