Jointly Extracting Interventions, Outcomes, and Findings from RCT Reports with LLMs

Somin Wadhwa WADHWA.S@NORTHEASTERN.EDU
Khoury College of Computer Sciences
Northeastern University
Boston, MA, USA

Jay DeYoung DEYOUNG.J@NORTHEASTERN.EDU
Khoury College of Computer Sciences
Northeastern University
Boston, MA, USA

Benjamin Nye BNYE@COLORADOCOLLEGE.EDU
Mathematics & Computer Science
Colorado College
Colorado Springs, CO, USA

Silvio Amir S.AMIR@NORTHEASTERN.EDU
Khoury College of Computer Sciences
Northeastern University
Boston, MA, USA

Byron C. Wallace B.WALLACE@NORTHEASTERN.EDU
Khoury College of Computer Sciences
Northeastern University
Boston, MA, USA

Abstract

Results from Randomized Controlled Trials (RCTs) establish the comparative effectiveness of interventions, and are in turn critical inputs for evidence-based care. However, results from RCTs are presented in (often unstructured) natural language articles describing the design, execution, and outcomes of trials; clinicians must manually extract findings pertaining to interventions and outcomes of interest from such articles. This onerous manual process has motivated work on (semi-)automating extraction of structured evidence from trial reports. In this work we propose and evaluate a text-to-text model built on instruction-tuned Large Language Models (LLMs) to jointly extract *Interventions*, *Outcomes*, and *Comparators* (ICO elements) from clinical abstracts, and infer the associated results reported. Manual (expert) and automated evaluations indicate that framing evidence extraction as a conditional generation task and fine-tuning LLMs for this purpose realizes considerable (∼20 point absolute F1 score) gains over the previous SOTA. We perform ablations and error analyses to assess aspects that contribute to model performance, and to highlight potential directions for further improvements. We apply our model to a collection of published RCTs through mid-2022, and release a searchable database of structured findings: http://ico-relations.ebm-nlp.com.

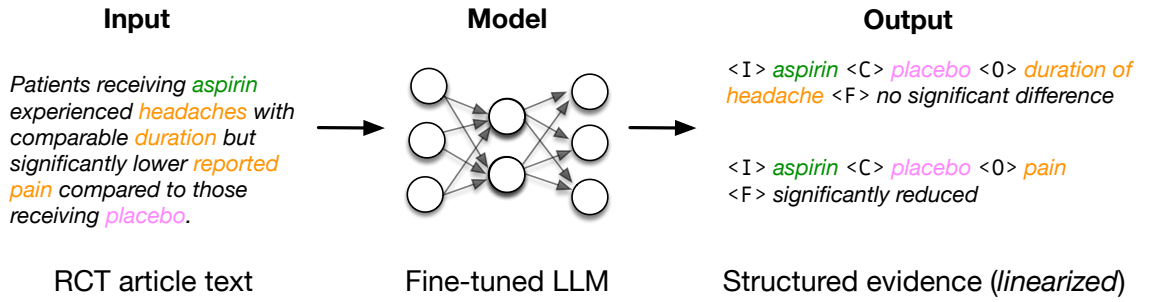

Figure 1: We fine-tune a Large Language Model (LLM) to map from free-text descriptions of clinical trials to structured representations of findings.

1. Introduction

Robust medical evidence concerning the comparative effectiveness of treatments is primarily disseminated in published free-text articles that report outcomes from randomized controlled trials (RCTs). Such trial results are critical inputs for practicing *Evidence-based medicine* (EBM; Sackett 1997), which seeks to inform patient care using the totality of relevant findings. Trial results are also potentially important for augmenting clinical predictions (Naik et al., 2022), and for calibrating trust in treatment suggestions offered by AI support systems (Yang et al., 2023), which ought to agree with the established evidence.

A challenge to making use of all available evidence is that findings from trials are disseminated via unstructured published articles. Researchers and healthcare providers must trawl through these to extract findings relevant to their clinical question(s). This problem has been exacerbated by the rapid production of new evidence: A now outdated estimate suggests that 75 trial reports are published *every single day* (Bastian et al., 2010); more recent estimates put this number at ∼140 trial reports per day (Marshall et al., 2020).

To allow practitioners to draw upon newly published evidence as it accumulates, we need tools that make navigating findings more efficient. This has motivated work on Natural Language Processing (NLP) methods to semi-automate aspects of data extraction from clinical trial reports (Kang et al. 2021; Kiritchenko et al. 2010; Wallace et al. 2016; Nye et al. 2022, *inter alia*). In this work we capitalize on and extend recent advances in NLP, specifically *instruction-tuned* LLM capabilities (Chung et al., 2022), to perform end-to-end structured evidence extraction from free-text (Figure 1). We achieve state-of-the-art (SOTA) performance on this challenging task: The model we introduce yields a ∼20 point absolute gain in F1 score over the prior SOTA approach. We ablate model components to assess their contributions. We also release model weights, and a database of structured findings inferred by our model over a comprehensive dataset of articles describing RCTs.

Generalizable Insights about Machine Learning in the Context of Healthcare

With respect to *healthcare*, this work makes significant progress on the important practical problem of structured evidence extraction from published articles describing RCTs. The outputs of this system may aid evidence synthesis, and might also serve as inputs to other

machine learning models in healthcare which could benefit from conditioning on robust evidence. Beyond this, the need for data extraction from free-text (e.g., clinical notes) is widespread in healthcare: Improved extraction methods have the potential to ultimately allow clinicians to focus on providing patient care instead of navigating unstructured data.

In terms of *machine learning*, we introduce and evaluate a method for training LLMs to perform a complex instance of *relation extraction*, a long-standing problem in ML (Ireson et al., 2005). To our knowledge, this is one of the first efforts to evaluate LLMs for medical relation extraction; we find that they outperform existing systems for this task by a large margin. As an additional contribution which may be of interest to the broader machine learning community, our ablations indicate that including *evidence spans* in extraction targets is an an important design decision—this complements recent developments inducing LLMs to provide free-text "rationales" for their outputs (Wei et al., 2022), and may have implications for those working with LLMs for relation extraction going forward.

2. Related Work

In this work we develop and evaluate methods using LLMs to extract results from clinical trial reports. Information and Relation Extraction (RE), generally, are well established sub-fields within NLP (Cowie and Lehnert, 1996), and we do not attempt to provide a general survey here. Instead, we contextualize our work by reviewing closely related efforts that focus on: (i) Information extraction from biomedical/clinical texts (Section 2.1); (ii) Models for jointly identifying entities and inferring relations between them (Section 2.2); and (iii) Recent approaches that treat RE as a *text-to-text* problem, a strategy that we adopt here (Section 2.3).

2.1. Information Extraction from Biomedical Literature and Clinical Text

A line of prior work in NLP attempts to extract relevant *Populations, Interventions, Comparators* and *Outcomes* (PICO elements) from clinical texts (Kim et al., 2011). Nye et al. (2018) collected a corpus of 5,000 annotated RCT abstracts and introduced novel NLP tasks aiding evidence-based medicine. Lee and Sun (2019) highlighted important aspects of PICO human-annotations to refine datasets by adopting a relaxed agreement schemes for human annotations of PICO. Jin and Szolovits (2018) introduced baselines in detecting PICO elements at the sentence level using LSTMs. Schmidt et al. (2020) proposed framing PICO extraction as a question-answering task and subsequently using transformer models, including SciBERT (Beltagy et al., 2019) — a masked language model pretrained on large-scale scientific data. These efforts either pre-dated Transformers, or used small encoder backbones, i.e., BERT (Devlin et al., 2018), rather than the generative models we use here.

Elsewhere, Lehman et al. (2019) introduced the *evidence inference* dataset which entailed inferring which medical treatments work with respect to a *given* ICO-set of interest. Using this dataset as a starting point, Nye et al. (2022) considered the end-to-end task of extracting PICO elements *and* inferring results (as opposed to performing inference for a given ICO triplet). They proposed an *extractive* entity extraction-linking-inference (ELI) sequential approach for this challenging task, and showed that it yielded results superior to standard joint architectures for relation extraction (Wadden et al., 2019). We improve upon

these earlier efforts by introducing an end-to-end *generative* model for the task of medical evidence inference.

2.2. Jointly Extracting Entities and their Relations

Early work in RE used pipeline approaches comprising separate models to, first, extract entities from a span of text, and then infer relations between those entities (if any). More recently, researchers have introduced joint extraction models since they tend to reduce error propagation and can capitalize on the connections between entities and their relations (Wang and Lu, 2020). Traditionally, such joint extraction methods principally worked by predicting "BILOU" tags (Beginning, Inside, Last, Outside, and Unit) for tokens in the input (Bekoulis et al., 2018b,a; Miwa and Bansal, 2016; Zheng et al., 2017; Verga et al., 2018). Span-based approaches extend these methods by constructing spans of tokens and then labeling these with respect to specific entity types, which enables processing of overlapping entities (Eberts and Ulges, 2019; Wadden et al., 2019) .

2.3. Generative Relation Extraction

Most earlier methods for identifying entities and extracting relations in free text trained models with a joint objective (Eberts and Ulges, 2021; Wang and Lu, 2020). The recent rise in (*very*) large language models (LLMs) (Brown et al., 2020; Chung et al., 2022) has motivated research into using these models for structured prediction tasks such as named entity recognition and RE (Nayak and Ng, 2019; Paolini et al., 2021; Huguet Cabot and Navigli, 2021; Wadhwa et al., 2023). This usually entails *linearizing*—that is, encoding into strings—the structured information and then tasking models with generating linearized target relations conditioned on corresponding inputs.

Building on these efforts, we propose to train and evaluate models to conditionally *generate* ICO spans, findings regarding the reported comparative effectiveness of the corresponding intervention compared to the comparator for the outcome in question, *and supporting textual evidence*. Specifically, we fine-tune an LLM to generate sets of linearized outputs (tuples) containing all the entities, relations, and supporting evidence from a given input RCT abstract (Figure 3).

3. Methods

3.1. End-to-End Evidence Inference

The task of *clinical evidence inference* comprises two sub-tasks: (i) Extraction of sets of relevant medical elements, i.e. ICO triplets; and (ii) Inference regarding the effect of the primary intervention on the outcome (i.e., *significant increase, significant decrease, no significant effect*), given the available evidence. These two subtasks can be seen as specialized instances of entity tagging and relation extraction, respectively. Recent work on clinical evidence inference has adopted a sequential (pipeline) approach in which ICO extraction is treated as a sequence tagging step, and then a separate inference module processes the tagged entities (Nye et al., 2022). This specialized approach outperformed model variants that attempted to jointly perform the task. However, prior methods for joint extraction and inference pre-dated the modern LLMs which are the current dominant paradigm in

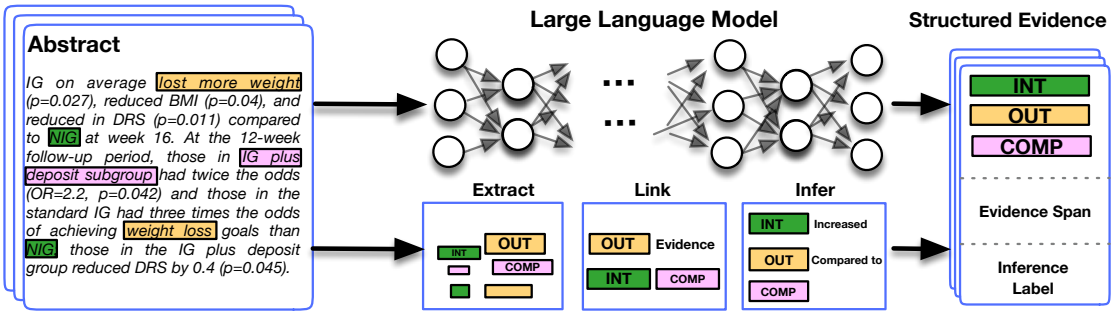

Figure 2: We propose instructional fine-tuning a large language model (top) using standard supervision to elicit evidence within generated ICO tuples. This approach yields substantial improvements over existing joint extraction approaches (bottom) where the entire task is decomposed into different *independent* phases.

NLP. Here we adopt such models, and treat the task of end-to-end evidence inference as a conditional language generation task (Figure 2).

Our targets are linearized strings comprising *multiple* tuples, each containing the elements (*Intervention*, *Comparator*, *Outcome*, *Evidence*, *Inference label*), extracted directly from an input abstract describing a RCT. Formally, given a RCT abstract $\mathcal{C}$, we model the probability of generating a linearized string y of length T containing N tuples (separated by special tokens in the linearized forms), conditioned on $\mathcal{C}$:

$$p_{\text{LM}}(y|\mathcal{C}) = \prod_{t=1}^{T} p(y_t|\mathcal{C}, y_{<t})$$

This is the standard (conditional) language modeling objective, and we optimize for per token cross-entropy loss. During training, we "teacher force", i.e., condition production of target token y_t on the reference sequence $y_{<t}$ and $\mathcal{C}$. At test time, the model iteratively conditions on its own outputs (we use greedy decoding).

The number of tuples associated with inputs is variable; language model flexibly models this by allowing the model to produce a special `EOS` token after enumerating all tuples. Note, however, that the model is unconstrained, and so can—and sometimes does, as we discussion in Section 4.2—produce invalid outputs (i.e., which do not conform to the linearized structured we assume).

Figure 3 provides an illustrative example where the abstract comprises two unique reference tuples:

`(zinc sulfate capsules, placebo, warts,` *`warts resolved in`* 68% *`of the patients in treatment group and`* 64% *`of the patients in placebo group`*`, no significant difference)`

`(zinc sulfate capsules, placebo, recurrence of warts,` *`three patients in treatment group and six patients in placebo group had a recurrence of warts (p=`*.19*`)`*`, no significant difference)`

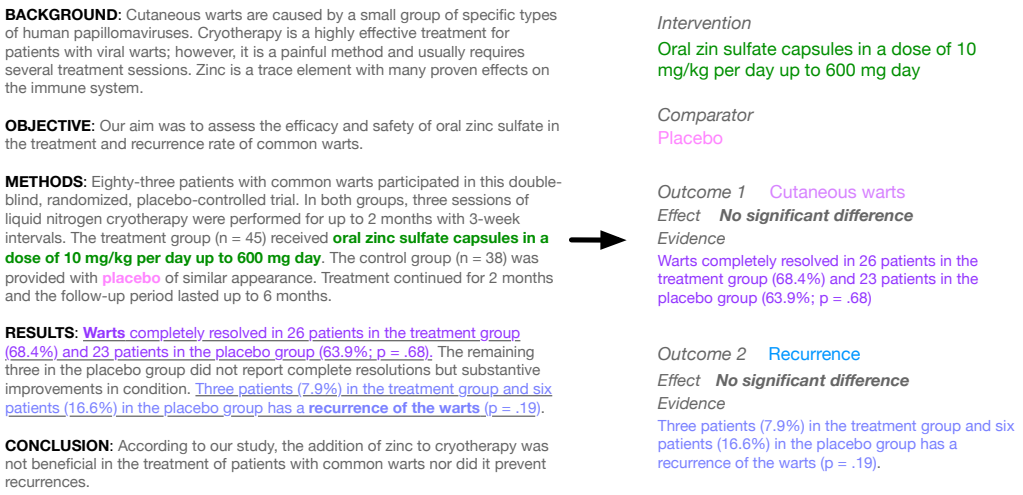

Figure 3: An illustration of the full evidence inference task. An end-to-end model is expected to extract all ICOs for which results were reported (highlighted here in pink, green, and orange) in an abstract describing an RCT, and infer a label (*significant increase*, *significant decrease*, *no significant difference*) based on the relevant evidence snippets which are also to be output (underlined here).

3.2. Data

We derived the data we use for training from the Evidence Inference dataset (Lehman et al., 2019; DeYoung et al., 2020). This comprises articles describing RCTs annotated by medical doctors.[1] An instance in this dataset comprises an abstract annotated with five elements: An ICO triplet, a *label* that indicates the directionality of a reported effect of the intervention for the given outcome relative to the comparator (i.e., categorizing that the intervention yielded *statistically significant increase, decrease, no effect* with respect to the outcome), and an *evidence snippet*. The latter is an excerpt from the abstract providing support for a particular label. This may be viewed as an explanation or "rationale". Together, these five elements form our targets. Table 1 provides basic data statistics for our training, validation, and test sets.

Evaluation Data To get an accurate assessment of model performance, Nye et al. (2022) also collected *exhaustive* manual annotations from medical experts for 160 RCT abstracts. Owing to the inherent noise in distantly-supervised training lables, we observed that human annotators often identify substantially more tuples per abstract — 4.97 tuples per abstract in the *validation* set, and 4.01 in the *test* set, as opposed to 2.76 in the (non-exhaustive) *training* set (Table 1). We provide more detailed examples of this phenomenon in our error analysis in Section 4.2.

1. Although the full dataset contains full-text RCT reports, here use use an abstract-only subset.

	Train		Dev		Test	
Abstracts	1,964	(1.00)	46	(1.00)	89	(1.00)
Total ICO Tuples	5,430	(2.76)	229	(4.97)	357	(4.01)
Unique ICO Triplets	4,951	(2.52)	224	(4.86)	351	(3.94)

Table 1: Dataset statistics. We report the number of abstracts and the number of relations per abstract (denoted parenthetically). Development and test set statistics differ from their source (Nye et al., 2022) as we omit documents with no annotated relations.

Full Inference End to End	Precision	Recall	F-1
BRAN (Verga et al., 2018)	0.05	0.41	0.08
DyGIE++ (Wadden et al., 2019)	0.24	0.13	0.17
ELI (Nye et al., 2022)	0.33	0.31	0.32
(end-to-end generation of ICO triplets with labels and supporting evidence)			
BART (Lewis et al., 2020)	0.38	0.33	0.35
T5-*base* (Raffel et al., 2020)	0.56	0.35	0.43
Flan-T5-*base* (Chung et al., 2022)	0.69	0.43	0.53
Flan-T5-*large*	**0.75**	**0.48**	**0.59**
Flan-T5-*large* (without evidence span extraction)	0.49	0.36	0.41

Table 2: End-to-end relation extraction results, compare to Nye et al. (2022) Table 2a

3.3. Experimental Setup

We performed all of our experiments on a single NVIDIA Quadro RTX 8000 GPU. We used the Huggingface library (v4.26.1; Wolf et al. 2020) and publicly available checkpoints.[2] of the language models we used in our experiments Our best performing model was trained for 8 epochs with a learning rate of $1e-6$, batch size of 2 (for both training and evaluation), with a maximum input length of 1024, and maximum output length of 512. For hyperparameter tuning, we only varied the learning rate, and max epochs. The remaining hyperparameters were left to their default values. We used the Adam optimizer without gradient accumulation or gradient checkpointing.

4. Results

We perform both an end-to-end evaluation (Table 2) and ablate performance over ICO-triplet extractions only (Table 3), maintaining comparability to existing work (Nye et al., 2022). Section 4.1 contains details of our manual evaluation, and Section 4.2 a detailed error analysis of model performance.

2. https://huggingface.co/docs/transformers/model_doc/flan-t5

ICO-Triplet Extraction	Precision	Recall	F-1
DyGIE++ (Wadden et al., 2019)	0.45	0.47	0.46
ELI (Nye et al., 2022)	0.46	0.69	0.55
(end-to-end generation of ICO-triplets)			
T5-base (Raffel et al., 2020)	0.68	0.62	0.65
Flan-T5-*base* (Chung et al., 2022)	0.78	0.68	0.73
Flan-T5-*large*	**0.85**	**0.74**	**0.79**

Table 3: ICO-Triplet Ablation, compare to Nye et al. (2022) Table 2b (entity extraction)

4.1. Evaluation

Open-ended free text generation poses challenges to the evaluation of model outputs. Past work in the area, especially prior to LLMs, tended to perform a "strict" evaluation (Taillé et al., 2020) requiring exact matches of entities and their corresponding relations to reference targets. This was appropriate because the models were effectively annotating input tokens, and references are assumed to be extractive. By contrast, because they are abstractive, LLMs can produce a variety of outputs that convey the desired semantic content—i.e., aligned with the reference target—without matching words exactly.

This motivates manual evaluation of RE outputs. Specifically, we recruited three medical doctors (domain experts) via the Upwork platform.[3] We asked these experts to individually evaluate each reference (to measure precision) *and* generated tuple (to measure recall) from our exhaustive test set. For each reference tuple we asked experts to indicate: (1) Whether the reference ICO triplet appears in the set of generated tuples for that given abstract; and (2) Whether the target tuple as a whole could be derived from the set of generated tuples for that given abstract. Similarly, for each generated tuple we asked annotators to indicate: (1) Whether the ICO triplet appears in the abstract; and (2) Whether the tuple as a whole is correct (i.e., if it also gets the relevant supporting evidence and reported directionality). We provide examples of each category in the Appendix A. Human evaluators achieved strong annotation agreement; Fleiss kappa, $\kappa = 0.77$. All three evaluators chose the same relevance label ∼92.4% of the time. We derived final (consensus) labels by simple majority vote.

4.2. Error Analysis

We now describe, and provide examples of, some of the recurring error types from our best performing model (Flan-T5-large) on the validation data, and a set of abstracts from approximately 660,000 RCTs from the Trialstreamer database.[4]

Incorrectly structured outputs The model sometimes generated incorrectly formatted outputs which cannot be evaluated because they do not conform to the expected structure. (Recall that the model is not explicitly constrained to yield outputs that follow the desired linearization scheme.) These include generations where: (1) there are missing elements in the (partial) ICO triplets; (2) outputs have an invalid syntactic structure (and are thus

3. https://upwork.com. We paid these experts $30/hour to evaluate generated tuples.
4. https://trialstreamer.ieai.robotreviewer.net/

unparseable by any downstream tools); (3) some elements are duplicated; (4) the output contains irrelevant or unrelated tokens. The following is an example of one such instance:

Generated: `[none, score, no, none, score was not significantly different between the two groups., no significant difference]`

Here the instance has an incorrect number of tuple elements (6 instead of 5), multiple elements are invalid, and while it does produce a valid label ("no significant difference"), there are no primary intervention and outcome spans associated with the label. This behavior occurs in only a small fraction (∼0.53%) of the RCT abstracts from Trialstreamer we ran through our model.

Opposite inference labels for same ICOs Approximately 12.3% of generated tuples had ICO-triplet matches in the reference set (i.e., the ICO triplet was correctly extracted), but the inferred label regarding the reported findings concerning these was incorrect (e.g., significant *increase* instead of significant *decrease*). On inspection we found that such tuples belonged to two categories: (1) The primary intervention and comparator were swapped (leading to a flipped, albeit still correct, inference label with the same extracted evidence span); (2) Minor differences in generated *outcomes* which resulted in a change in the label. The following is an example of the latter from our development set (PMID: 24227660:[5])

Abstract snippet: `Canagliflozin increased urinary glucose excretion in a dose-dependent manner and produced statistically significant reductions in body weight compared with placebo (least squares mean percent changes from baseline of -2.2%, -2.9%, -2.7%, and -1.3% with canagliflozin 50, 100, and 300 mg and placebo; P < 0.05 for all comparisons). Overall adverse event (AE) rates were similar across groups. Canagliflozin was associated with higher rates of genital mycotic infections in women, which were generally mild and led to few study discontinuations. Osmotic diuresis-related AE rates were low and similar across groups.`

Reference: `[canagliflozin, body weight, placebo, Canagliflozin increased urinary glucose excretion in a dose-dependent manner and produced statistically significant reductions in body weight compared with placebo., canagliflozin [LABEL] significantly decreased [OUT] body weight [COMP] placebo]`

Generated: `[canagliflozin, body weight reduction, placebo, Canagliflozin increased urinary glucose excretion in a dose-dependent manner and produced statistically significant reductions in body weight compared with placebo., canagliflozin [LABEL] significantly increased [OUT] body weight reduction [COMP] placebo]`

An *increase* in *body weight reduction* is functionally the same as a *decrease* in *body weight*, and this explains the label flip.

5. https://pubmed.ncbi.nlm.nih.gov/24227660/

Combining multiple tuples On average, our best performing model generates 3.49 ICO tuples per instance, as opposed to 4.01 per instance in the reference test set (Table 1). This difference appears to be due to the model *combining* multiple interventions and/or outcomes into one in cases where the inference label is preserved, in turn reducing the number of generated tuples. Consider the following example[6] from our dev set where this behavior can be observed (PMID: 27981024[7]):

Reference: `[memory game with fruit, banana intake, no fruit game, evidence, significant increase], [memory game with fruit, mandarin intake, no fruit game, evidence, significant increase]`

Generated: `[fruit version of memory game, intake of mandarins and bananas, no fruit game, evidence, significant increase]`

Here we can observe that the generated tuple has combined banana and mandarin intake, yielding a single output instead of the two in the reference.

Correctly generated but without any corresponding reference This type of "error" is limited to non-exhaustive reference sets, and occurs when there is no corresponding reference tuple for a correctly generated ICO output (because the reference set is non-exhaustive). While this is rare, instances featuring this type of error highlight the utility of the retrieved evidence snippets, which can be used to verify the output. The following is one such example taken from an abstract that was **not** exhaustively annotated (PMID: 28211020[8]):

Abstract snippet:[9] `High-risk patients undergoing brain surgery were randomly assigned to a usual care group (control group) or a GDFR group. In the GDFR group, (1) fluid maintenance was restricted to 3 ml/kg/h of a crystalloid solution and (2) colloid boluses were allowed only in case of hypotension associated with a low cardiac index and a high stroke volume variation. The primary outcome variable was ICU length of stay, and secondary outcomes were lactates at the end of surgery, postoperative complications, hospital length of stay, mortality at day 30, and costs. ICU length of stay was shorter (3 days [1-5] vs. 6 days [3-11], p = 0.001) and ICU costs were lower in the GDFR group. The total number of complications (46 vs. 99, p = 0.043) and the proportion of patients who developed one or more complications (19.2 vs. 34.7%, p = 0.034) were smaller in the GDFR group. Hospital length of stay and costs, as well as mortality at 30 day, were not significantly reduced.`

Reference: `[goal-directed fluid restriction (GDFR) strategy, median hospital length of stay, usual care, *Hospital length of stay and costs, as well as mortality at 30-day, were not significantly reduced*, no significant difference]`

6. Example simplified for brevity.
7. https://pubmed.ncbi.nlm.nih.gov/27981024/
8. https://pubmed.ncbi.nlm.nih.gov/28211020/
9. Shortened for brevity.

Generated: [goal-directed fluid restriction (GDFR), ICU length of stay, usual care group (control group), *ICU length of stay was shorter (3 days [1{5] vs. 6 days [3{11], p = 0.001) and ICU costs were lower in the GDFR group.*, significantly decreased]

Here, on inspecting only the ICO-triplet and the inference label, one might assume that the generated tuple is incorrect with respect to the reference (due to a flipped label). However, examining the extracted evidence in support of the label and the full abstract reveals that the study does indeed report *median length of hospital stay* and *ICU-length of stay* as separate outcomes with different (opposite) labels.

5. A Prototype for Browsing Structured Evidence

To further demonstrate the (potential) utility of structured evidence extraction over the published evidence base, we make available a demonstration web application.[10] This permits free-text search, which retrieves relevant structured evidence extracted from papers (we also link back to the original PubMed articles).

We processed all Randomized Control Trials indexed by Trialstreamer (Marshall et al., 2020) as of June 2022, yielding 657,698 total studies and a total of 1,204,027 extracted relations. Relation extraction required 584 GPU (32GB NVIDIA V100) hours. Of the 770,356 unique Trialstreamer documents, approximately 50k instances were missing a full abstract. When processed via FLAN, 74k (about 10%) had an unparseable output; lacking (or possessing an extra) a syntatic element (e.g. missing a bracket or having an extra one, or other terminator symbol). Another 5k had an output with an incorrect number of fields. 82 had a malformed label. When parsing misclassified RCTs (erroneously included in Trialstreamer), the model would hallucinate ICOs and findings not present in the data.

The prototype implements a BM25 search (Robertson et al., 1994) backed by SQLite (Hipp, 2020), allowing for search over multiple fields.[11] The website allows for downloading search results (by search or by list of PMIDs/PMCIDs); our hope is that this may be of interest to researchers. We will make the entire raw database of inferred relations available upon publication.

6. Discussion

We have introduced and evaluated a state-of-the-art approach to end-to-end structured evidence extraction from natural language articles describing the conduct and results of clinical trials. Specifically, we treat this problem as a conditional generation task and fine-tune Flan-T5 (Chung et al., 2022)—a modestly sized instruction-tuned sequence-to-sequence model—to consume unstructured texts and yield structured tuples composed of interventions, comparators, outcomes and the results reported regarding these. The latter comprises a discrete prediction encoding the direction of the reported finding, and a snippet of evidence supporting this determination. Ablations indicate the importance of jointly extracting evidence spans to support the inference task; this may have implications for work on relation extraction via conditional generative models more broadly.

10. Hosted at http://ico-relations.ebm-nlp.com.
11. We experimented with embedding based methods but were ultimately disappointed with results

Figure 4: A screenshot of our prototype search interface over structured evidence. (A) User inputs a search query and select the fields (B) to be searched over via an SQL search (C; Hipp 2020, e.g., entire abstract, only ICOs). Search results can either be downloaded as a structured CSV (D) or the user can browse through individual results (E). We retrieve up to 100 documents per search query with 10 documents per page (F). The interface allows the users to read expanded abstracts, view structured findings *(shown above)*, and expand structured markup for a tabular view of findings.

Structured evidence extraction is an important task for realizing the promise of evidence-based medicine (EBM; Sackett 1997), which aspires to inform treatment decisions on the basis of all available relevant evidence. The vast (unstructured) evidence base and rapid accumulation of new findings render practicing EBM challenging. The proposed approach to evidence extraction achieves substantially better performance than the prior state-of-the-art (Nye et al., 2022), and this brings us closer to being able to synthesize all evidence relevant to a given query, in real-time.

To illustrate the potential utility of this model, we have also made available a prototype interface that permits search directly over structured evidence tuples automatically extracted from a comprehensive database of randomized controlled trial reports. Our hope is that this demonstrates the precision of model outputs, and suggests how such extracted evidence might help researchers and healthcare providers navigate the evidence base more efficiently than is currently possible. We also anticipate that the resultant database (comprising tuples from all RCTs in humans) may be a useful resource for researchers in machine learning for healthcare broadly, as one might draw upon such trial results to inform and/or justify ML predictions (Yang et al., 2023; Naik et al., 2022).

Limitations This work has several important limitations. First, while we have reported promising empirical results, the model we have trained here still makes errors (e.g., provides inexhaustive extractions from an inputs; see Section 4.2). Any downstream use of the structured evidence outputs need to take this into account.

A methodological limitation is that we did not investigate the capabilities of even larger LLMs like GPT-3.5/4 Brown et al. (2020) for this task. One could, in principle, use OpenAI's API to fine-tune such models for this task, and given their size it is likely that this would yield (probably moderately) improved results. We opted not to pursue this primarily because we prefer to use open-source models, to ensure scientific transparency and so that we can release model weights. Furthermore, the main contribution here is the framing of the task as a language modeling problem; the particular choice of underlying LLM is a secondary consideration.

Finally, while we think structured evidence in the format that we have extracted—providing explicit sets of interventions, comparators, outcomes and evidence concerning these—will provide meaningful downstream utility for those interested in navigating and making sense of the published evidence base, it is currently an intermediate output. The actual utility of this sort of model for downstream tasks which ultimately might affect care will require conducting further research.

Acknowledgments

This work was supported by the National Institutes of Health (NIH) under award R01LM012086, and by the National Science Foundation (NSF) award 1750978.

References

Hilda Bastian, Paul Glasziou, and Iain Chalmers. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS medicine*, 7(9):e1000326, 2010.

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Systems with Applications*, 114:34–45, dec 2018a. doi: 10.1016/j.eswa.2018.07.032. URL https://doi.org/10.1016%2Fj.eswa.2018.07.032.

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. Adversarial training for multi-context joint entity and relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2830–2836, Brussels, Belgium, October-November 2018b. Association for Computational Linguistics. doi: 10.18653/v1/D18-1307. URL https://aclanthology.org/D18-1307.

Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1371. URL https://aclanthology.org/D19-1371.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022.

Jim Cowie and Wendy Lehnert. Information extraction. *Communications of the ACM*, 39 (1):80–91, 1996.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Jay DeYoung, Eric Lehman, Benjamin Nye, Iain Marshall, and Byron C. Wallace. Evidence inference 2.0: More data, better models. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 123–132, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.bionlp-1.13. URL https://aclanthology.org/2020.bionlp-1.13.

Markus Eberts and Adrian Ulges. Span-based joint entity and relation extraction with transformer pre-training. *CoRR*, abs/1909.07755, 2019. URL http://arxiv.org/abs/1909.07755.

Markus Eberts and Adrian Ulges. An end-to-end model for entity-level relation extraction using multi-instance learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3650–3660, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.319. URL https://aclanthology.org/2021.eacl-main.319.

Richard D Hipp. SQLite, 2020. URL https://www.sqlite.org/index.html.

Pere-Lluís Huguet Cabot and Roberto Navigli. REBEL: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.204. URL https://aclanthology.org/2021.findings-emnlp.204.

Neil Ireson, Fabio Ciravegna, Mary Elaine Califf, Dayne Freitag, Nicholas Kushmerick, and Alberto Lavelli. Evaluating machine learning for information extraction. In *Proceedings of the 22nd international conference on Machine learning*, pages 345–352, 2005.

Di Jin and Peter Szolovits. PICO element detection in medical text via long short-term memory neural networks. In *Proceedings of the BioNLP 2018 workshop*, pages 67–75, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-2308. URL https://aclanthology.org/W18-2308.

Tian Kang, Ali Turfah, Jaehyun Kim, Adler Perotte, and Chunhua Weng. A neuro-symbolic method for understanding free-text medical evidence. *Journal of the American Medical Informatics Association*, 28(8):1703–1711, 2021.

Su Nam Kim, David Martinez, Lawrence Cavedon, and Lars Yencken. Automatic classification of sentences to support evidence based medicine. In *BMC bioinformatics*, volume 12, pages 1–10. BioMed Central, 2011.

Svetlana Kiritchenko, Berry de Bruijn, Simona Carini, Joel Martin, and Ida Sim. ExaCT: automatic extraction of clinical trial characteristics from journal publications. *BMC medical informatics and decision making*, 10(1):56, 2010.

Grace E. Lee and Aixin Sun. A study on agreement in pico span annotations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 1149–1152, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361729. doi: 10.1145/3331184.3331352. URL https://doi.org/10.1145/3331184.3331352.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. Inferring which medical treatments work from reports of clinical trials. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3705–3717,

Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1371. URL https://aclanthology.org/N19-1371.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703.

Iain J Marshall, Benjamin Nye, Joël Kuiper, Anna Noel-Storr, Rachel Marshall, Rory Maclean, Frank Soboczenski, Ani Nenkova, James Thomas, and Byron C Wallace. Trialstreamer: A living, automatically updated database of clinical trial reports. *Journal of the American Medical Informatics Association*, 27(12):1903–1912, 2020.

Makoto Miwa and Mohit Bansal. End-to-end relation extraction using LSTMs on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1116, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1105. URL https://aclanthology.org/P16-1105.

Aakanksha Naik, Sravanthi Parasa, Sergey Feldman, Lucy Lu Wang, and Tom Hope. Literature-augmented clinical outcome prediction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 438–453, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.33. URL https://aclanthology.org/2022.findings-naacl.33.

Tapas Nayak and Hwee Tou Ng. Effective modeling of encoder-decoder architecture for joint entity and relation extraction, 2019.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1019. URL https://aclanthology.org/P18-1019.

Benjamin E. Nye, Jay DeYoung, Eric Lehman, Ani Nenkova, Iain J. Marshall, and Byron C. Wallace. Understanding clinical trial reports: Extracting medical entities and their relations, 2022.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, RISHITA ANUBHAI, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. Structured prediction as translation between augmented natural languages. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=US-TP-xnXI.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2020.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at trec-3. In *Text Retrieval Conference*, 1994.

David L Sackett. Evidence-based medicine. In *Seminars in perinatology*, volume 21, pages 3–5. Elsevier, 1997.

Lena Schmidt, Julie Weeds, and Julian P. T. Higgins. Data mining in clinical trial text: Transformers for classification and question answering tasks. In *International Conference on Health Informatics*, 2020.

Bruno Taillé, Vincent Guigue, Geoffrey Scoutheeten, and Patrick Gallinari. Let's Stop Incorrect Comparisons in End-to-end Relation Extraction! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3689–3701, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.301. URL https://aclanthology.org/2020.emnlp-main.301.

Patrick Verga, Emma Strubell, and Andrew McCallum. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 872–884, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1080. URL https://aclanthology.org/N18-1080.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5784–5789, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1585. URL https://aclanthology.org/D19-1585.

Somin Wadhwa, Silvio Amir, and Byron Wallace. Revisiting relation extraction in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589, Toronto, Canada, July 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.acl-long.868.

Byron C Wallace, Joël Kuiper, Aakash Sharma, Mingxi Zhu, and Iain J Marshall. Extracting pico sentences from clinical trial reports using supervised distant supervision. *The Journal of Machine Learning Research*, 17(1):4572–4596, 2016.

Jue Wang and Wei Lu. Two are better than one: Joint entity and relation extraction with table-sequence encoders. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1706–1721, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.133. URL https://aclanthology.org/2020.emnlp-main.133.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

Qian Yang, Yuexing Hao, Kexin Quan, Stephen Yang, Yiran Zhao, Volodymyr Kuleshov, and Fei Wang. Harnessing biomedical literature to calibrate clinicians' trust in ai decision support systems. In *Conference on Human Factors in Computing Systems (CHI)*, 2023.

Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, and Bo Xu. Joint extraction of entities and relations based on a novel tagging scheme. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1227–1236, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1113. URL https://aclanthology.org/P17-1113.

Appendix A: Examples

Here we provide some full length output examples generated directly from our best performing models. Abstracts have been shortened for brevity to include key parts.

PMID 1457358

Abstract Micronuclei reflect DNA damage in exfoliated cells and may thus provide a marker of early-stage carcinogenesis. Pre-treatment blood levels of cotinine, beta-carotene, retinol and vitamins C and E were similar in the placebo group (n = 61) and the treatment group (n = 53). Plasma beta-carotene levels increased 13-fold in the treatment group during intervention. Initial micronuclei counts (per 3,000 cells) were higher in the treatment group than in the placebo group (5.0 vs 4.0, P ¡ 0.05). During intervention, the treatment group showed a 47% decrease, whereas the placebo group showed a non-significant decrease (16%). After adjustment for the initial levels, the treatment group had 27% lower micronuclei counts than the placebo group at the end of the trial (95% CI: 9-41%). These results indicate that beta-carotene may reduce lung cancer risk in man by preventing DNA damage in early-stage carcinogenesis.

Reference [*beta-carotene supplementation (20 mg d-1), Initial micronuclei counts (per 3,000 cells), placebo, Initial micro nuclei counts (per 3,000 cells) were higher in the treatment group than in the placebo group (5.0 vs 4.0, P; 0.05).,* **[INT]** *beta-carotene supplementation (20 mg d-1)* **[LABEL]** *significantly increased* **[OUT]** *Initial micronuclei counts (per 3,000 cells)* **[COMP]** *placebo*]

Generated [*14 weeks of beta-carotene supplementation (20 mg d-1), micronuclei counts, placebo, Initial micronuclei counts (per 3,000 cells) were higher in the treatment group than in the placebo group (5.0 vs 4.0, P 0.05).,* **[INT]** *14 weeks of beta-carotene supplementation (20 mg d-1)* **[LABEL]** *significantly increased* **[OUT]** *micronuclei counts* **[COMP]** *placebo*]

PMID 29295869

Abstract Since sCD163 is shed to serum by inflammatory signals including lipopolysaccharides (LPS, endotoxin), we investigated sCD163 and correlations with lipid metabolism following LPS exposure. Eight healthy male subjects were investigated on two separate occasions: (i) following an LPS exposure and (ii) following saline exposure. Each study day consisted of a four-hour non-insulin-stimulated period followed by a two-hour hyperinsulinemic euglycemic clamp period. A 3H-palmitate tracer was used to calculate the rate of appearance (Rapalmitate). Blood samples were consecutively obtained throughout each study day. Abdominal subcutaneous adipose tissue was obtained for western blotting. We observed a significant two-fold increase in plasma sCD163 levels following LPS exposure (P ¡ 0.001), and sCD163 concentrations correlated positively with the plasma concentration of free fatty acids, Rapalmitate, lipid oxidation rates and phosphorylation of the hormone-sensitive lipase at serine 660 in adipose tissue (P

< 0.05, all). Furthermore, sCD163 concentrations correlated positively with plasma concentrations of cortisol, glucagon, tumour necrosis factor (TNF)-α, interleukin (IL)-6 and IL-10 ($P < 0.05$, all). We observed a strong correlation between sCD163 and stimulation of lipolysis and fat oxidation following LPS exposure. These findings support preexisting theory that inflammation and macrophage activation play a significant role in lipid metabolic adaptions under conditions such as obesity, DM2 and NAFLD.

Reference [*LPS exposure, macrophage activation, Saline axposure, We observed a significant two-fold increase in plasma sCD163 levels following LPS exposure ($P < 0.001$), and sCD163 concentrations correlated positively with the plasma concentration of free fatty acids, Rapalmitate, lipid oxidation rates and phosphorylation of the hormone-sensitive lipase at serine 660 in adipose tissue ($P < 0.05$, all). Furthermore, sCD163 concentrations correlated positively with plasma concentrations of cortisol, glucagon, tumour necrosis factor (TNF)-α, interleukin (IL)-6 and IL-10 ($P < 0.05$, all).*, **[INT]** *LPS exposure* **[LABEL]** *significantly increased* **[OUT]** *macrophage activation* **[COMP]** *Saline axposure*]

Generated [*lipopolysaccharides (LPS) exposure, plasma sCD163 levels, saline exposure, We observed a significant two-fold increase in plasma sCD163 levels following LPS exposure ($P < 0.001$), and sCD163 concentrations correlated positively with the plasma concentration of free fatty acids, Rapalmitate, lipid oxidation rates and phosphorylation of the hormone-sensitive lipase at serine 660 in adipose tissue ($P < 0.05$, all).*, **[INT]** *lipopolysaccharides (LPS) exposure* **[LABEL]** *significantly increased* **[OUT]** *plasma sCD163 levels* **[COMP]** *saline exposure*]

PMID 26258157

Abstract We conducted a double-blind, randomized trial of 134 outpatients with polymerase chain reaction-confirmed influenza to assess the effects of oseltamivir initiated 48–119 hours after illness onset. Oseltamivir treatment did not reduce illness duration, severity, or duration of virus detection. However, the power of this study was limited due to lower than expected enrollment.

Reference 1 [*Oseltamivir, Severity of illness, Control, Oseltamivir treatment did not reduce illness duration, severity, or duration of virus detection.*, **[INT]** *Oseltamivir* **[LABEL]** *no significant difference* **[OUT]** *Severity of illness* **[COMP]** *Control*]

Reference 2 [*Oseltamivir, Duration of virus detection, Control, Oseltamivir treatment did not reduce illness duration, severity, or duration of virus detection.*, **[INT]** *Oseltamivir* **[LABEL]** *no significant difference* **[OUT]** *Duration of virus detection* **[COMP]** *Control*]

Generated [*Oseltamivir, Illness duration, severity, or duration of virus detection, Control, Oseltamivir treatment did not reduce illness duration, severity, or duration of virus detection.*, **[INT]** *Oseltamivir* **[LABEL]** *no significant difference* **[OUT]** *Illness duration, severity, or duration of virus detection* **[COMP]** *Control*]

PMID 26283840

Abstract Most commonly practiced surgical "lay open" technique to treat fistula-in-ano (a common anorectal pathology) has high rate of recurrence and anal incontinence. Alternatively, a nonsurgical cost efficient treatment with Ksharasutra (cotton Seton coated with Ayurvedic medicines) has minimal complications. In our study, we have tried to compare these two techniques. A prospective randomized control study was designed involving patients referred to the Department of General Surgery in RG Kar Medical College, Kolkata, India, from January 2010 to September 2011. Among 50 patients, 26 were in Ksharasutra and 24 were in fistulotomy group. 86% patients were male and 54% of the patients were in the fourth decade. About 74% fistulas are inter-sphincteric and 26% were of trans-sphincteric variety. Severe postoperative pain was more (7.7% vs. 25%) in fistulotomy group, while wound discharge was more associated with Ksharasutra group (15.3% vs. 8.3%). Wound scarring, bleeding, and infection rate were similar in both groups. Ksharasutra group took more time to heal (mean: 53 vs. 35.7 days, P = 0.002) despite reduced disruption to their routine work (2.7 vs. 15.5 days work off, P < 0.001). Interestingly, pain experienced was less in Ksharasutra group, there was no open wound in contrast to fistulotomy and it was significantly cost effective (Rupees 166 vs 464).

Reference [*Ksharasutra (cotton Seton coated with Ayurvedic medicines), number of days "off-work", fistulotomy, Ksharasutra group took more time to heal (mean: 53 vs 35.7 days, P = 0.002) despite reduced disruption to their routine work (2.7 vs. 15.5 days work off, P < 0.001).,* **[INT]** *Ksharasutra (cotton Seton coated with Ayurvedic medicines)* **[LABEL]** *significantly decreased* **[OUT]** *number of days "off-work"* **[COMP]** *fistulotomy*]

Generated 1 [*Ksharasutra (cotton Seton coated with Ayurvedic medicines), Severe postoperative pain, fistulotomy, Severe postoperative pain was more (7.7% vs. 25%) in fistulotomy group, while wound discharge was more associated with Ksharasutra group (15.3% vs. 8.3%).,* **[INT]** *Ksharasutra (cotton Seton coated with Ayurvedic medicines)* **[LABEL]** *significantly decreased* **[OUT]** *Severe postoperative pain* **[COMP]** *fistulotomy*]

Generated 2 [*Ksharasutra (cotton Seton coated with Ayurvedic medicines), Wound scarring, bleeding, and infection rate, fistulotomy, Wound scarring, bleeding, and infection rate were similar in both groups.,* **[INT]** *Ksharasutra (cotton Seton coated with Ayurvedic medicines)* **[LABEL]** *no significant difference* **[OUT]** *Wound scarring, bleeding, and infection rate* **[COMP]** *fistulotomy*]