Fact or Fiction: Verifying Scientific Claims

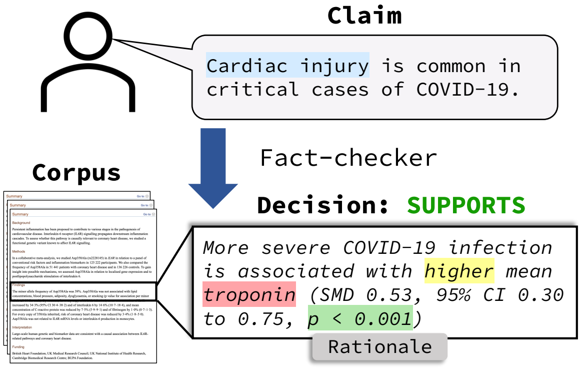
David Wadden[†*] **Shanchuan Lin**[†] **Kyle Lo**[‡] **Lucy Lu Wang**[‡]
Madeleine van Zuylen[‡] **Arman Cohan**[‡] **Hannaneh Hajishirzi**[†‡]
[†] University of Washington, Seattle, WA, USA
[‡] Allen Institute for Artificial Intelligence, Seattle, WA, USA
{dwadden,linsh,hannaneh}@cs.washington.edu
{kylel,lucyw,madeleinev,armanc}@allenai.org

Abstract

We introduce scientific claim verification, a new task to select abstracts from the research literature containing evidence that SUPPORTS or REFUTES a given scientific claim, and to identify rationales justifying each decision. To study this task, we construct SCIFACT, a dataset of 1.4K expert-written scientific claims paired with evidence-containing abstracts annotated with labels and rationales. We develop baseline models for SCIFACT, and demonstrate that simple domain adaptation techniques substantially improve performance compared to models trained on Wikipedia or political news. We show that our system is able to verify claims related to COVID-19 by identifying evidence from the CORD-19 corpus. Our experiments indicate that SCIFACT will provide a challenging testbed for the development of new systems designed to retrieve and reason over corpora containing specialized domain knowledge. Data and code for this new task are publicly available at https://github.com/allenai/scifact. A leaderboard and COVID-19 fact-checking demo are available at https://scifact.apps.allenai.org.

1 Introduction

Due to rapid growth in the scientific literature, it is difficult for researchers – and the general public even more so – to stay up to date on the latest findings. This challenge is especially acute during public health crises like the current COVID-19 pandemic, due to the extremely fast rate at which new findings are reported and the risks associated with making decisions based on outdated or incomplete information. As a result, there is a need for automated tools to assist researchers and the public in evaluating the veracity of scientific claims.



Cardiac injury is common in critical cases of COVID-19.

More severe COVID-19 infection is associated with higher mean troponin (SMD 0.53, 95% CI 0.30 to 0.75, p < 0.001)

Figure 1: A scientific claim, supported by evidence identified by our system. To correctly verify this claim, the system must possess background knowledge that *troponin* is a protein found in cardiac muscle and that elevated levels of *troponin* are a marker of *cardiac injury*. In addition, it must be able to reason about directional relationships between scientific processes: replacing *higher* with *lower* would cause the rationale to REFUTE the claim rather than SUPPORT it. Finally, the system should interpret $p < 0.001$ as an indication that the reported finding is statistically significant.

Fact-checking – a task in which the veracity of an input *claim* is verified against a corpus of documents that *support* or *refute* the claim – has been studied to combat the proliferation of misinformation in political news, social media, and on the web (Thorne et al., 2018; Hanselowski et al., 2019). However, verifying scientific claims poses new challenges to both dataset construction and effective modeling. While political claims are readily available on fact-checking websites and can be verified by crowd workers, annotators with extensive domain knowledge are required to generate and verify scientific claims.

In addition, NLP systems for scientific claim verification must possess additional capabilities beyond those required to verify factoid claims. For instance, to verify the claim shown in Figure 1, a

[*] Work performed during internship with the Allen Institute for Artificial Intelligence.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 7534–7550,
November 16–20, 2020. ©2020 Association for Computational Linguistics

Claim 1: Lopinavir / ritonavir have exhibited favorable clinical responses when used as a treatment for coronavirus.
Supports: . . . *Interestingly, after lopinavir/ritonavir (Kaletra, AbbVie) was administered, β-coronavirus viral loads significantly decreased and no or little coronavirus titers were observed.*
Refutes: *The focused drug repurposing of known approved drugs (such as lopinavir/ritonavir) has been reported failed for curing SARS-CoV-2 infected patients.* It is urgent to generate new chemical entities against this virus . . .

Claim 2: The coronavirus cannot thrive in warmer climates.
Supports: *...most outbreaks display a pattern of clustering in relatively cool and dry areas...This is because the environment can mediate human-to-human transmission of SARS-CoV-2, and unsuitable climates can cause the virus to destabilize quickly...*
Refutes: *...significant cases in the coming months are likely to occur in more humid (warmer) climates, irrespective of the climate-dependence of transmission and that summer temperatures will not substrantially limit pandemic growth*.

Table 1: Evidence identified by our system as supporting and refuting two claims concerning COVID-19.

system must have the ability to access scientific background knowledge, reason over increases and decreases in quantities or measurements, and make sense of specialized statistical language.

In this paper, we introduce the task of scientific claim verification to evaluate the veracity of scientific claims against a scientific corpus. Table 1 presents some examples. To facilitate research on this task, we construct SciFact, an expert-annotated dataset of 1,409 scientific claims accompanied by abstracts that support or refute each claim, and annotated with rationales (Lei et al., 2016) justifying each Supports / Refutes decision. To create the dataset, we develop a novel annotation protocol in which annotators re-formulate naturally occurring claims in the scientific literature – *citation sentences* – into atomic scientific claims. Using citation sentences as a source of claims both speeds the claim generation process and guarantees that the topics discussed in SciFact are representative of the research literature. In addition, citation links indicate the exact documents likely to contain evidence necessary to verify a given claim.

We establish performance baselines on SciFact with an approach similar to DeYoung et al. (2020a), which achieves strong performance on the Fever claim verification dataset (Thorne et al., 2018). Our baseline is a pipeline system which retrieves abstracts related to an input claim, uses a BERT-based (Devlin et al., 2019) sentence selector to identify rationale sentences, and labels each abstract as Supports, Refutes, or NoInfo with respect to the claim. We demonstrate that our baseline can benefit from training on claims from domains including Wikipedia articles and politics.

We showcase the ability of our model to verify expert-written claims concerning the novel coronavirus COVID-19 against the newly-released CORD-19 corpus (Wang et al., 2020). Expert annotators judge retrieved evidence to be plausible for 23 of 36 claims.[1] Our results and analyses demonstrate the importance of the new task and dataset to support significant future research in this domain.

In summary, our contributions include: (1) We introduce and formalize the scientific claim verification task. (2) We develop a novel annotation protocol to generate and verify 1.4K naturally-occurring claims about scientific findings. (3) We establish strong baselines on this task, and identify substantial opportunities for improvement at all stages of the modeling pipeline. (4) We demonstrate the efficacy of our system in a real-world case study verifying claims about COVID-19 against the research literature.

2 Background and task definition

As illustrated in Figure 1, scientific claim verification is the task of identifying evidence from the research literature that Supports or Refutes a given scientific claim. Table 1 shows the results of our system applied to claims about the novel coronavirus COVID-19. For each claim, the system identifies relevant scientific abstracts, and labels the relation of each abstract to the claim as either Supports or Refutes. Verifying scientific claims is challenging and requires domain-specific background knowledge – for instance, in order to identify the evidence supporting Claim 1 in Table 1, the system must determine that a reduction in coronavirus viral load indicates a favorable clinical response, even though this fact is never mentioned.

Scientific claims In SciFact, a scientific claim is an *atomic verifiable statement* expressing a finding

[1] We emphasize that our model is a *research prototype* and should not be used to make any medical decisions whatsoever.

about one aspect of a scientific entity or process, which can be verified from a single source.[2] For instance, "*The* R_0 *of the novel coronavirus is 2.5*" is valid, but opinion-based statements like "*The government should require people to stand six feet apart to stop coronavirus*" are not. Compound claims like "*Aerosolized coronavirus droplets can travel at least 6 feet and can remain in the air for 3 hours*" should be split into two atomic claims.

Claims in SciFact are *natural* – they are derived from citation sentences, or *citances* (Nakov et al., 2004), that occur naturally in scientific articles. This is similar to political fact-checking datasets such as UKP Snopes (Hanselowski et al., 2019), which use political fact-checking websites as a source of natural claims. On the other hand, claims in the popular FEVER dataset (Thorne et al., 2018) are *synthetic*, since they are created by annotators by mutating sentences from the Wikipedia articles that will serve as evidence.

Supporting and refuting evidence In most fact-checking work, claims are assigned a global truth label based on the entirety of the available evidence. For example in FEVER, the claim "*Barack Obama was the 44th President of the United States*" can be verified using Wikipedia as an evidence source.

While SciFact claims are indeed verifiable assertions about scientific findings, accurately assigning a global truth label to a scientific claim (given a fixed scientific corpus) requires a systematic review by a team of experts. In this work we focus on the simpler task of assigning SUPPORTS or REFUTES relations to individual *claim-abstract pairs*.

Each SUPPORTS or REFUTES relation between claim and abstract must be justified by at least one *rationale*. A rationale is a minimal collection of sentences which, taken together as premises in the context of the abstract, can reasonably be judged by a domain expert as implying the claim. Rationales facilitate the development of *interpretable models* which not only have the ability to make label predictions, but can also identify the exact sentences that are necessary for their decisions.

3 The SciFact dataset

The SciFact dataset consists of 1,409 scientific claims[3] verified against a corpus of 5,183 abstracts. Abstracts that support or refute each claim are annotated with rationales. We describe our corpus creation and annotation process.

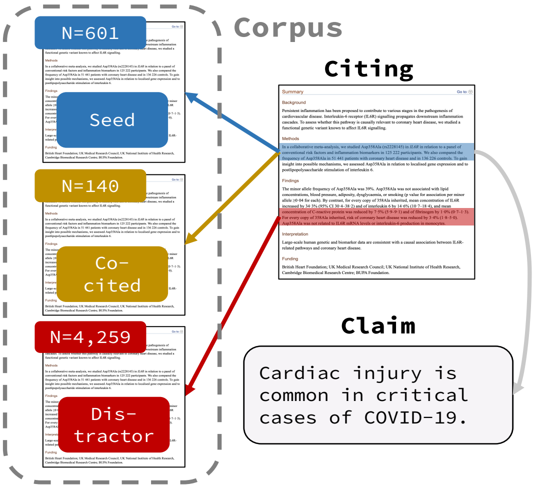


Figure 2: Corpus construction. Citing abstracts are identified for each seed document. A claim is written based on the source citance in the citing abstract.

3.1 Data source and corpus construction

To construct SciFact, we use S2ORC (Lo et al., 2020), a publicly-available corpus of millions of scientific articles. To ensure that documents in our dataset are of high quality, we randomly sample articles from a manually curated collection of well-regarded journals spanning domains from basic science (e.g., *Cell*, *Nature*) to clinical medicine (e.g., *JAMA*, *BMJ*). The full list of journals is included in Appendix C.1. We restrict to articles with at least 10 citations. The resulting collection is referred to as our *seed* set. We use the S2ORC citation graph to sample *source citances* from *citing articles* which cite these seed articles. If a citance cites other articles not in the seed set, we refer to these as *co-cited* articles and add them to the corpus, as depicted in Figure 2. The content of the cited abstracts encompasses a diverse array of topics within biomedicine, as shown in Figure 3. The majority of citances used for SciFact cite only the seed article (no co-cited articles), as we found in initial annotation experiments that these citances tended to yield specific, easy-to-verify claims.

To expand the corpus, we identify five papers cited in the same paper as each source citance but in a different paragraph, and add these to the corpus as *distractor abstracts*. These abstracts often

[2] Requiring annotators to search multiple sources increases cognitive burden and decreases annotation quality.

[3] SciFact is comparable in size to recent scientific datasets for tasks such as QA (e.g. PubMedQA (Jin et al., 2019) has 1,000 questions), and information extraction (e.g. SciERC (Luan et al., 2018) has 500 annotated abstracts).

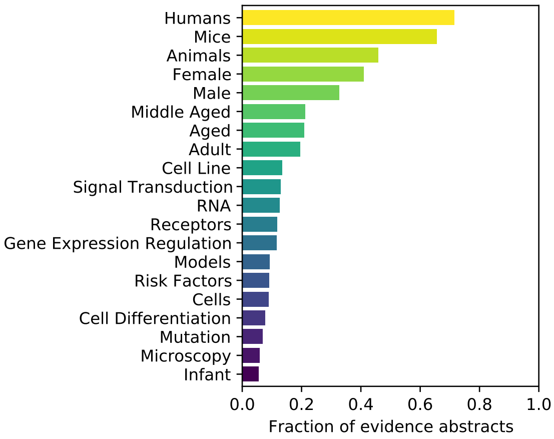


Figure 3: Most frequently occurring Medical Subject Headings (MeSH) terms (y-axis) among cited abstracts. MeSH is a controlled vocabulary used for indexing articles in PubMed. Topics range from clinical trial reports ("Humans", "Risk Factors") to molecular biology ("Cell Line", "RNA").

discuss similar topics to the evidence documents, increasing the difficulty of abstract retrieval and making our metrics more accurately reflect the system's performance on a large research corpus.

3.2 Claim writing

Annotation Annotators are shown a source citance in the context of an article, and are asked to write up to three claims based on the content of the citance; see Appendix C.2 for an example. This results in *natural* claims because the annotator does not see the cited article's abstract – the *cited abstract* – at the time of claim writing. Annotators are asked to skip citances that do not make statements about specific scientific findings.

The claim writers included four experts with background in scientific NLP, fifteen undergraduates studying the life sciences, and four graduate students (doctoral or medical) in the life sciences. Detailed information on the annotator training process can be found in Appendix C.3. The claim-writing interface is shown in Appendix D.

Claim negation Unless the authors of the source citance were mistaken, cited articles should provide supporting evidence for the claims made in a citance. To obtain examples where an abstract REFUTES a claim, an NLP expert wrote negations of existing claims, taking precautions not to bias the negations by using obvious keywords like "not" (Schuster et al., 2019; Gururangan et al., 2018). In §6.1, we demonstrate that a "claim-only" verification model performs poorly, suggesting that the negation process did not introduce severe artifacts.

3.3 Claim verification

Annotation For each claim, all of the claim's cited abstracts are annotated for evidence. Annotators are shown a single claim - cited abstract pair, and asked to label the pair as SUPPORTS, REFUTES, or NOINFO. Although our task definition allows for a single claim to be both supported and refuted (by different abstracts) – an occurrence we observe on real-world COVID-19 claims (§6.3) – this never occurs in our dataset. Each claim has a single label. Counts for each label are shown in Table 2a. Overall, the annotators found evidence in 63% of cited abstracts. If the annotator assigns a SUPPORTS or REFUTES label, they must also identify all rationales as defined in §2. Table 2b provides statistics on the number of sentences per rationale, the number of rationales per claim / abstract pair, and the number of evidence abstracts per claim. No abstract has more than 3 rationales for a given claim, and all rationales consist of at most three sentences. Rationales in SCIFACT are mutually exclusive. 28 rationales contain non-contiguous sentences.

The verifiers included three NLP experts, five life science undergraduates, and five graduate students studying life sciences. Annotators verified claims that they did not write themselves. Annotation guidelines are provided in Appendix D.

SCIFACT claims are verified against abstracts rather than full articles since (1) abstracts can be annotated more scalably, (2) evidence is found in the abstract in more than 60% of cases, and (3) previous attempts at full-document annotation suffered from low annotator agreement (§7).

Quality We assign 232 claim-abstract pairs for independent re-annotation. The label agreement is 0.75 Cohen's κ, comparable with the 0.68 Fleiss' κ reported in Thorne et al. (2018), and 0.70 Cohen's κ reported in Hanselowski et al. (2019). To measure rationale agreement, we treat each sentence as either classified as "part of a rationale" or "not part of a rationale" and compute sentence-level agreement. The resulting Cohen's κ is 0.71.

4 The SCIFACT task

Task Formulation The inputs to our task are a scientific claim c and a corpus of abstracts $\mathcal{A}$. All abstracts $a \in \mathcal{A}$ are labeled as $y(c, a) \in \{$SUPPORTS, REFUTES, NOINFO $\}$ with respect to a claim c.

Fold	Supports	NoInfo	Refutes	All
Train	332	304	173	809
Dev	124	112	64	300
Test	100	100	100	300
All	556	516	337	1409

(a) Distribution of claim labels in SciFact.

	0	1	2	3+
Cited abstracts per claim	-	1278	86	45
Evidence abstracts per claim	516	830	37	26
Rationales per abstract	-	552	290	153
Sentences per rationale	-	1542	92	11

(b) Evidence counts at various levels of granularity. For example, Column 2 of the row "Rationales / abstract" indicates that 290 claim / abstract pairs are supported by 2 distinct rationales.

Table 2: Statistics on claim labels, and the number of evidence abstracts and rationales per claim.

The abstracts that either Support or Refute c are referred to as *evidence abstracts* for c, denoted as $\mathcal{E}(c)$. Each evidence abstract $a \in \mathcal{E}(c)$ is annotated with rationales. A single rationale R_i is a collection of sentences $\{r_1(c, a), \ldots, r_m(c, a)\}$, where m is the number of sentences in rationale R_i. We denote the set of all rationales as $\mathcal{R}(c, a) = \{R_1(c, a), \ldots, R_n(c, a)\}$.

Given a claim c and a corpus $\mathcal{A}$, the system must predict a set of evidence abstracts $\widehat{\mathcal{E}}(c)$. For each abstract $a \in \widehat{\mathcal{E}}(c)$, it must predict a label $\widehat{y}(c, a)$, and a collection of *rationale sentences* $\widehat{S}(c, a) = \{\widehat{s}_1(c, a), \ldots, \widehat{s}_\ell(c, a)\}$. Note that although the gold annotations may contain multiple separate rationales, to simplify the prediction task we only require the model to predict a single collection of *rationale sentences*; these sentences may encompass multiple gold rationales.

Task Evaluation We evaluate the task at two levels of granularity. For *abstract-level* evaluation, we assess the model's ability to identify the abstracts that support or refute the claim. For *sentence-level* evaluation, we evaluate the model's performance at identifying the sentences sufficient to justify the abstract-level predictions. We conduct evaluations in both the "Open" Fever-style (Thorne et al., 2018) setting where the evidence abstracts must be retrieved, and the "Oracle abstract" ERASER-style (DeYoung et al., 2020a) setting where the gold evidence abstracts $\mathcal{E}(c)$ are provided.

Abstract-level evaluation is inspired by the Fever score. Given a claim c, a predicted evidence abstract $a \in \widehat{\mathcal{E}}(c)$ is *correctly labeled* if (1) a is a gold evidence abstract for c, and (2) The predicted label is correct: $\widehat{y}(c, a) = y(c, a)$. It is *correctly rationalized* if, in addition, the predicted rationale sentences contain a gold rationale, i.e., there exists some gold rationale $R_i(c, a) \subseteq \widehat{S}(c, a)$.

Like Fever, which limits the maximum number of predicted rationale sentences to five, SciFact limits to three predicted rationale sentences. Overall performance is measured by the micro-F1 of the precision and recall over the correctly-labeled and correctly-rationalized evidence abstracts. We refer to these evaluations as $\text{Abstract}_{\text{Label-Only}}$ and $\text{Abstract}_{\text{Label+Rationale}}$, respectively.

Sentence-level evaluation measures performance in identifying individual rationale sentences. Unlike the abstract-level metrics, this evaluation penalizes the prediction of extra rationale sentences.

A predicted rationale sentence $\widehat{s}(c, a)$ is correctly *selected* if (1) It is a member of some gold rationale $R_i(c, a)$, (2) all other sentences from the same gold rationale $R_i(c, a)$ are among the predicted $\widehat{S}(c, a)$, and (3) $\widehat{y}(c, a) \neq$ NoInfo[4]. It is correctly *labeled* if, in addition, the abstract a is correctly labeled: $\widehat{y}(c, a) = y(c, a)$.

Overall performance is measured by the micro-F1 of the precision and recall of correctly-selected and correctly-labeled rationale sentences, denoted $\text{Sentence}_{\text{Selection-Only}}$ and $\text{Sentence}_{\text{Selection+Label}}$. For sentence-level evaluation, we do not limit the number of predicted rationale sentences, since the evaluation penalizes models that over-predict.

5 VeriSci: Baseline model

We develop a baseline (referred to as VeriSci) that takes a claim c and corpus $\mathcal{A}$ as input, identifies evidence abstracts $\widehat{\mathcal{E}}(c)$, and predicts a label $\widehat{y}(c, a)$ and rationale sentences $\widehat{S}(c, a)$ for each $a \in \widehat{\mathcal{E}}(c)$. Following the "BERT-to-BERT" model presented in DeYoung et al. (2020a); Soleimani et al. (2019), VeriSci is a pipeline of three components:

1. AbstractRetrieval retrieves k abstracts with highest TF-IDF similarity to the claim.
2. RationaleSelection identifies rationale sentences $\widehat{S}(c, a)$ for each abstract.
3. LabelPrediction makes the final label prediction $\widehat{y}(c, a)$.

Rationale selection Given a claim c and abstract a, we train a model to predict $z_i \triangleq$

[4] Condition (3) eliminates rationale sentences which were identified by the rationale selector, but proved insufficient to justify a final Supports / Refutes decision

$\mathbb{1}[a_i \text{ is a rationale sentence}]$ for each sentence a_i in a. For each sentence, we encode the concatenated sequence $w_i = [a_i, \text{SEP}, c]$ using a BERT-style language model and predict a score $\tilde{z}_i = \sigma[f(\text{CLS}(w_i))]$, where σ is the sigmoid function, f is a linear layer and $\text{CLS}(w_i)$ is the CLS token from the encoding of w_i. We train the model on pairs of claims and their cited abstracts and minimize cross-entropy loss between z_i and $\tilde{z}_i$. For each claim, we use cited abstracts labeled NoInfo, as well as non-rationale sentences from abstracts labeled Supports and Refutes as negative examples. To make predictions, we select all sentences a_i with $\tilde{z}_i > t$ as rationale sentences, where $t \in [0, 1]$ is tuned on the dev set (Appendix A.1).

Label prediction Sentences identified by the rationale selector are passed to a separate BERT-based model to make the final labeling decision. Given a claim c and abstract a, we concatenate the claim and the predicted rationale sentences $u = [\widehat{s}_1(c, a), \dots \widehat{s}_\ell(c, a), \text{SEP}, c]$[5], and predict $\tilde{y}(c, a) = \phi[f(\text{CLS}(u))]$, where ϕ is the softmax function, and f is a linear layer with three outputs representing the {Supports, Refutes, NoInfo } labels. We minimize the cross-entropy loss between $\tilde{y}(c, a)$ and the true label $y(c, a)$.

We train the model on pairs of claims and their cited abstracts using gold rationales as input. For cited abstracts labeled NoInfo, we choose the k sentences from the cited abstract with highest TF-IDF similarity to the claim as input rationales. For prediction, we use the predicted rationale sentences $\widehat{S}(c, a)$ as input and predict $\hat{y}(c, a) = \text{argmax}\ \tilde{y}(c, a)$. NoInfo is predicted for abstracts with no rationale sentences.

We experimented with a label prediction model which encodes entire abstracts via the Longformer (Beltagy et al., 2020), and makes predictions using the document-level CLS token. Performance was not competitive with our pipeline setup, likely because the label predictor struggles to identify relevant information when given full abstracts.

6 Experiments

In our experiments, we (1) analyze the performance of each individual component of VeriSci, (2) evaluate full task performance in both the "Oracle abstract" and "Open" settings, (3) present promising results verifying claims about COVID-19 using VeriSci, and (4) discuss some modeling challenges presented by the dataset.

	Rational-Select.			Label-Pred.
Training data	P	R	F1	Acc.
Fever	41.5	57.9	48.4	67.6
UKP Snopes	42.5	62.3	50.5	71.3
SciFact	73.7	70.5	**72.1**	75.7
Fever + SciFact	72.4	67.2	69.7	**81.9**
Sentence encoder	P	R	F1	Acc.
SciBERT	74.5	74.3	**74.4**	69.2
BioMedRoBERTa	75.3	69.9	72.5	71.7
RoBERTa-base	76.1	66.1	70.8	62.9
RoBERTa-large	73.7	70.5	72.1	**75.7**
Model inputs	P	R	F1	Acc.
Claim-only	-	-	-	44.5
Abstract-only	60.1	60.9	60.5	53.3

Table 3: Comparison of different training datasets, encoders, and model inputs for RationaleSelection and LabelPrediction, evaluated on the SciFact dev set. The claim-only model cannot select rationales.

6.1 Pipeline components

We examine the effects of different training datasets, sentence encoders, and model inputs on the performance of the RationaleSelection and LabelPrediction modules. The RationaleSelection module is evaluated on its ability to select rationale sentences given gold abstracts[6]. The LabelPrediction module is evaluated on its 3-way label classification accuracy given gold rationales from cited abstracts. Cited abstracts labeled NoInfo are included in the evaluation. These abstracts have no gold rationale sentences; as in §5, we provide the k most similar sentences from the abstract as input (more details in Appendix A).

Training Data We train on (1) Fever, (2) UKP Snopes, (3) SciFact, and (4) Fever pretraining followed by SciFact fine-tuning. RoBERTa-large (Liu et al., 2019) is used as the sentence encoder.

Sentence encoder We fine-tune SciBERT (Beltagy et al., 2019), BioMedRoBERTa (Gururangan et al., 2020), RoBERTa-base, and RoBERTa-large. SciFact is used as training data.

Model Inputs We examine the performance of "claim-only" and "abstract-only" models trained on SciFact, using RoBERTa-large as the sentence encoder. The claim-only model makes label predic-

[5] We truncate the rationale input if it exceeds the BERT token limit. c is never truncated.

[6] Our Fever-trained RationaleSelection module achieves 79.9 sentence-level F1 on the Fever test set, virtually identical to 79.6 reported in DeYoung et al. (2020a).

			Sentence-level						Abstract-level					
			Selection-Only			Selection+Label			Label-Only			Label+Rationale		
Retrieval	Model		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Oracle abstract	Oracle rationale	1	100.0	80.5	$89.2_{2.1}$	89.6	72.2	$79.9_{3.0}$	90.1	77.5	$83.3_{2.4}$	90.1	77.5	$83.3_{2.4}$
	Zero-shot	2	42.5	45.1	$43.8_{2.0}$	36.1	38.4	$37.2_{2.3}$	86.9	53.6	$66.3_{3.1}$	67.9	41.9	$51.8_{3.4}$
	VERISCI	3	76.1	63.8	$69.4_{2.6}$	66.5	55.7	$\mathbf{60.6}_{3.1}$	87.3	65.3	$74.7_{2.8}$	84.9	63.5	$\mathbf{72.7}_{2.9}$
Open	Oracle rationale	4	100.0	56.5	$72.2_{3.3}$	87.6	49.5	$63.2_{3.7}$	88.9	54.1	$67.2_{3.2}$	88.9	54.1	$67.2_{3.2}$
	Zero-shot	5	28.7	37.6	$32.5_{2.3}$	23.7	31.1	$26.9_{2.3}$	56.0	42.3	$48.2_{3.3}$	42.3	32.0	$36.4_{3.3}$
	VERISCI	6	45.0	47.3	$46.1_{3.0}$	38.6	40.5	$\mathbf{39.5}_{3.0}$	47.5	47.3	$47.4_{3.1}$	46.6	46.4	$\mathbf{46.5}_{3.1}$

Table 4: Test set performance on SCIFACT, according to the metrics from §4. For the "Oracle abstract" rows, the system is provided with gold evidence abstracts. "Oracle rationale" rows indicate that the gold rationales are provided as input. "Zero-shot" indicates zero-shot performance of a verification system trained on FEVER. Additionally, standard deviations are reported as subscripts for all F1 scores. See Appendix B for standard deviations on all reported metrics.

tions based on the claim text alone, without access to evidence abstracts. The abstract-only model selects rationale sentences and makes label predictions without access to the claim.

Results The results are shown in Table 3. For LABELPREDICTION, the best performance is achieved by training first on the large FEVER dataset and then fine-tuning on the smaller in-domain SCIFACT training set. To understand the benefits of FEVER pretraining, we examined the claim / evidence pairs where the FEVER + SCIFACT- trained model made correct predictions but the SCIFACT- trained model did not. In 36 / 44 of these cases, the SCIFACT-trained model predicts NOINFO. Thus pretraining on FEVER appears to improve the model's ability to recognize textual entailment relationships between evidence and claim – particularly relationships indicated by non-domain-specific cues like "is associated with" or "has an important role in".

For RATIONALESELECTION, training on SCIFACT alone produces the best results. We examined the rationales that the SCIFACT- trained model identified but the FEVER- trained model missed, and found that they generally contain science-specific vocabulary. Thus, training on additional out-of-domain data provides little benefit.

RoBERTa-large exhibits the strongest performance on label prediction, while SCIBERT has a slight edge on rationale selection. The "claim-only" model exhibits very poor performance, which provides some reassurance that the claim negation procedure described in §3.2 does not introduce obvious statistical artifacts. Similarly, the poor performance of the "abstract-only" model indicates that the model needs access to the claim being verified in order to identify relevant evidence.

6.2 Full task

Experimental setup Based on the results from §6.1, we use the RATIONALESELECTION module trained on SCIFACT only, and the LABELPREDICTION module trained on FEVER + SCIFACT for our final end-to-end system VERISCI. Although SCIBERT performs slightly better on rationale selection, using RoBERTa-large for both RATIONALESELECTION and LABELPREDICTION gave the best full-pipeline performance on the dev set, so we use RoBERTa-large for both components. For the ABSTRACTRETRIEVAL module, the best dev set full-pipeline performance was achieved by retrieving the top $k = 3$ documents.

Model comparisons We report performance of three model variants. For the "Oracle rationale" setting, the RATIONALESELECTION module is replaced by an oracle which outputs gold rationales for correctly retrieved documents, and no rationales for incorrect retrievals. The "Zero-shot" setting reports the zero-shot generalization performance of a model trained on FEVER (the results on UKP Snopes were slightly worse). VERISCI reports the performance of our best system.

Results The results are shown in Table 4. In the oracle abstract setting, the abstract-level F1 scores are roughly comparable to label classification accuracies, and the $\text{Abstract}_{\text{Label+Rationale}}$ score in Row 3 implies an end-to-end classification accuracy of roughly 70%, given gold abstracts.

Access to in-domain data during training clearly improves performance. Despite the small size of SCIFACT, training on these data

Reasoning type	Example	
Science background	**Claim:**	Rapamycin slows aging in fruit flies.
	Evidence:	*. . . feeding rapamycin to adult Drosophila produces life span extension . . .*
	Gold Verdict:	SUPPORTS
	Reasoning:	Drosophila is a type of fruit fly.
Directionality	**Claim:**	Inhibiting glucose-6-phospate dehydrogenase impairs lipogenesis
	Evidence:	*. . . suppression of 6PGD increased lipogenesis*
	Gold Verdict:	REFUTES
	Reasoning:	A decrease (not increase) in lipogenesis would indicate lipogenesis impairment.
Numerical reasoning	**Claim:**	Bariatric surgery improves resolution of diabetes.
	Evidence:	*Strong associations were found between bariatric surgery and the resolution of T2DM, with a HR of 9.29 (95% CI 6.84-12.62)...*
	Gold Verdict:	SUPPORTS
	Reasoning:	A HR (hazard ratio) that is greater than 1 with 95% confidence indicates improvement.
Cause and effect	**Claim:**	Major vault protein (MVP) functions to decrease tumor aggression.
	Evidence:	*Knockout of MVP leads to miR-193a accumulation...inhibiting tumor progression*
	Gold Verdict:	REFUTES
	Reasoning:	Knocking out (removing) MVP inhibits tumor progression → MVP increases tumor aggression.
Coreference	**Claim:**	Low saturated fat diets have adverse effects on the development of infants
	Evidence:	*Neurological development of children in the intervention group was at least as good as ... the control group*
	Gold Verdict:	REFUTES
	Reasoning:	The intervention group in this study was placed on a low saturated fat diet.

Table 5: Reasoning types required to verify SCIFACT claims which are classified incorrectly by our modeling baseline. Words crucial for correct verification are highlighted.

leads to relative improvements of 47% on open Sentence$_{\text{Selection+Label}}$, and 28% on open Abstract$_{\text{Label+Rationale}}$ over FEVER alone (Row 6 vs. Row 5). The three pipeline components make similar contributions to the overall model error. Replacing RATIONALESELECTION with an oracle leads to a roughly 20-point rise in Sentence$_{\text{Selection+Label}}$ F1 (Row 6 vs. Row 4). Replacing ABSTRACTRETRIEVAL with an oracle as well leads to a gain of roughly 20 more points (Row 4 vs. Row 1).

Nearly all correctly-labeled abstracts are supported by at least one rationale. There is only a two-point difference in F1 between Abstract$_{\text{Label-Only}}$ and Abstract$_{\text{Label+Rationale}}$ in the oracle setting (Row 3), and a one-point difference in the open setting (Row 6). The differences between Sentence$_{\text{Selection-Only}}$ and Sentence$_{\text{Selection+Label}}$ are larger, caused by examples where the model finds the evidence but fails to predict its relationship to the claim. We examine these in §6.4.

We evaluate the statistical robustness of our results by generating 10,000 bootstrap-resampled versions of the test set (Dror et al., 2018) and computing the standard deviation of all performance metrics. Table 4 shows the standard deviations in F1 score. Uncertainties on all metrics for both the dev and test set can be found in Appendix B. The results indicate that the observed differences in model performance are statistically robust and cannot be attributed to random variation in the dataset.

6.3 Verifying claims about COVID-19

We conduct exploratory experiments using our system to verify claims concerning COVID-19. We tasked a medical student to write 36 COVID-related claims. For each claim c, we used VERISCI to predict evidence abstracts $\widehat{\mathcal{E}}(c)$. The annotator examined each $(c, \widehat{\mathcal{E}}(c))$ pair. A pair was labeled *plausible* if $\widehat{\mathcal{E}}(c)$ was nonempty, and at least half of the evidence abstracts in $\widehat{\mathcal{E}}(c)$ were judged to have reasonable rationales and labels. For 23 / 36 claims, the response of VERISCI was deemed plausible by our annotator, demonstrating that VERISCI is able to successfully retrieve and classify evidence in many cases. Two examples are shown in Table 1. In both cases, our system identifies both supporting *and* refuting evidence.

6.4 Error analysis

To better understand the errors made by VERISCI, we conduct a manual analysis of test set predictions where an evidence abstract was correctly retrieved, but where the model failed to identify any relevant rationales or predicted an incorrect label. We identify five modeling capabilities required to correct these mistakes (Table 5 provides examples):

Science background includes knowledge of domain-specific lexical relationships.

Directionality requires understanding increases or decreases in scientific quantities.

Numerical reasoning involves interpreting numerical or statistical findings.

Cause and effect requires reasoning about counterfactuals.

Coreference involves drawing conclusions using context stated outside of a rationale sentence.

7 Related work

Fact checking and rationalized NLP models Fact-checking datasets include PolitiFact (Vlachos and Riedel, 2014), Emergent (Ferreira and Vlachos, 2016), LIAR (Wang, 2017), SemEval 2017 Task 8 RumorEval (Derczynski et al., 2017), Snopes (Popat et al., 2017), CLEF-2018 CheckThat! (Barrón-Cedeño et al., 2018), Verify (Baly et al., 2018), Perspectrum (Chen et al., 2019), FEVER (Thorne et al., 2018), and UKP Snopes (Hanselowski et al., 2019). Hanselowski et al. (2019) provides a thorough review. To our knowledge, there are no existing data sets for scientific claim verification. We refer to our task as "claim verification" rather than "fact-checking" to emphasize that our focus is to help researchers make sense of scientific findings, not to counter disinformation.

Fact-checking is one of a number of tasks where a model is required to justify a prediction via *rationales* from the source document. The ERASER dataset (DeYoung et al., 2020a) provides a suite of benchmark datasets (including SCIFACT) for evaluating rationalized NLP models.

Related scientific NLP tasks The *citation contextualization* task (Cohan et al., 2015; Jaidka et al., 2017) is to identify spans in a cited document that are relevant to a particular citation in a citing document. Unlike SCIFACT, these citations are not re-written into atomic claims and are therefore more difficult to verify. Expert annotators achieved very low (21.7%) inter-annotator agreement on the BioMedSumm dataset (Cohen et al., 2014), which contains 314 citations referencing 20 papers.

Biomedical question answering datasets include BioASQ (Tsatsaronis et al., 2015) and PubMedQA (Jin et al., 2019), which contain 855 and 1,000 "*yes / no*" questions respectively (Gu et al., 2020). Claim verification and question answering are both knowledge intensive tasks which require an understanding of the relationship between an input query and relevant supporting text.

Automated evidence synthesis (Marshall and Wallace, 2019; Beller et al., 2018; Tsafnat et al., 2014; Marshall et al., 2017) seeks to automate the process of creating *systematic reviews* of the medical literature[7] – for instance, by extracting PICO snippets (Nye et al., 2018) and inferring the outcomes of clinical trials (Lehman et al., 2019; DeYoung et al., 2020b). We hope that systems for claim verification will serve as components in future evidence synthesis frameworks.

8 Conclusion and future work

Claim verification allows us to trace the sources and measure the veracity of scientific claims. These abilities have emerged as particularly important in the context of the current pandemic, and the broader reproducibility crisis in science. In this article, we formalize the task of scientific claim verification, and release a dataset (SCIFACT) and models (VERISCI) to support work on this task. Our results indicate that it is possible to train models for scientific fact-checking and deploy them with reasonable efficacy on real-world claims related to COVID-19.

Scientific claim verification presents a number of promising avenues for research on models capable of incorporating background information, reasoning about scientific processes, and assessing the strength and provenance of various evidence sources. This last challenge will be especially crucial for future work that seeks to verify scientific claims against sources other than the research literature – for instance, social media and the news. We hope that the resources presented in this paper encourage future research on these important challenges, and help facilitate progress toward the broader goal of scientific document understanding.

Acknowledgments

This research was supported by the ONR MURI N00014-18-1-2670, ONR N00014-18-1-2826, DARPA N66001-19-2-4031, NSF (IIS 1616112), Allen Distinguished Investigator Award, and the Sloan fellowship. We thank the Semantic Scholar team at AI2, UW-NLP, and H2lab at UW for helpful comments and feedback.

[7] https://www.cochranelibrary.com/about/about-cochrane-reviews

References

Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018. Integrating stance detection and fact checking in a unified corpus. In *NAACL*.

Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez i Villodre, Pepa Atanasova, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. 2018. Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. task 2: Factuality. In *CLEF*.

Elaine Beller, Justin Clark, Guy Tsafnat, Clive Elliott Adams, Heinz Diehl, Hans Lund, Mourad Ouzzani, Kristina Thayer, James Thomas, Tari Turner, J. S. Xia, Karen A. Robinson, and Paul P Glasziou. 2018. Making progress with the automation of systematic reviews: principles of the international collaboration for the automation of systematic reviews (icasr). *Systematic Reviews*, 7.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *EMNLP*.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *ArXiv*, abs/2004.05150.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in nlp. In *EMNLP*.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle: Discovering diverse perspectives about claims. In *NAACL*.

Arman Cohan, Luca Soldaini, and Nazli Goharian. 2015. Matching citation text and cited spans in biomedical literature: a search-oriented approach. In *NAACL*.

Kevin Bretonnel Cohen, Hoa Trang Dang, Anita de Waard, Prabha Yadav, and Lucy Vanderwende. 2014. Tac 2014 biomedical summarization track. `https://tac.nist.gov/2014/BiomedSumm/`.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *SemEval*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020a. Eraser: A benchmark to evaluate rationalized nlp models. In *ACL*.

Jay DeYoung, Eric Lehman, Ben Nye, Iain James Marshall, and Byron C. Wallace. 2020b. Evidence inference 2.0: More data, better models. In *BioNLP@ACL*.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *ACL*.

Bradley Efron and Robert Tibshirani. 1993. An introduction to the bootstrap.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *NAACL*.

Yu Gu, Robert Tinn, Hao Cheng, M. Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. *ArXiv*, abs/2007.15779.

Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *ACL*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *NAACL*.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *CoNLL*.

Kokil Jaidka, Muthu Kumar Chandrasekaran, Devanshu Jain, and Min-Yen Kan. 2017. The cl-scisumm shared task 2017: Results and key insights. In *BIRNDL@JCDL*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *EMNLP*.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. 2019. Inferring which medical treatments work from reports of clinical trials. In *NAACL*.

Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2016. Rationalizing neural predictions. In *ACL*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S. Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *ACL*.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *EMNLP*.

Iain James Marshall, Joël Kuiper, Edward Banner, and Byron C. Wallace. 2017. Automating biomedical evidence synthesis: Robotreviewer. *ACL*.

Iain James Marshall and Byron C. Wallace. 2019. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic Reviews*, 8.

Preslav I Nakov, Ariel S Schwartz, and Marti Hearst. 2004. Citances: Citation sentences for semantic analysis of bioscience text. In *SIGIR workshop on Search and Discovery in Bioinformatics*.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain James Marshall, Ani Nenkova, and Byron C. Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *ACL*.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *WWW*.

Tal Schuster, Darsh J. Shah, Yun Jie Serene Yeo, Daniel Filizzola, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. In *EMNLP*.

Amir Soleimani, Christof Monz, and Marcel Worring. 2019. Bert for evidence retrieval and claim verification. In *European Conference on Information Retrieval*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *NAACL*.

Guy Tsafnat, Paul P Glasziou, Miew Keen Choong, Adam G. Dunn, Filippo Galgani, and Enrico W. Coiera. 2014. Systematic review automation technologies. *Systematic Reviews*, 3:74 – 74.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artières, Axel-Cyrille Ngonga Ngomo, Norman Heino, Éric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. In *BMC Bioinformatics*.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *ACL Workshop on Language Technologies and Computational Social Science*.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William. Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. Cord-19: The covid-19 open research dataset. *ArXiv*, abs/2004.10706.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *ACL*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

A Model implementation details

All models are implemented using the Huggingface Transformers package (Wolf et al., 2019).

A.1 Parameters for the final VERISCI system

For the ABSTRACTRETRIEVAL module, VERISCI retrieves the top $k = 3$ documents ranked by TF-IDF similarity using unigram + bigram features. These parameters are tuned on the SCIFACT development set.

When making predictions using the RATIONALESELECTION module described in §5, we find that the usual decision rule of predicting $\hat{z}_i = 1$ when $\tilde{z}_i \geq 0.5$ works well for models trained on SCIFACT. However, for models trained on FEVER and UKP Snopes, we achieve better performance by tuning the classification threshold t, such that $\hat{z}_i = 1$ when $\tilde{z}_i \geq t$, on the SCIFACT dev set. The best threshold was $t = 0.025$ when training on FEVER, and $t = 0.75$ when training on UKP Snopes.

A.2 Training the RATIONALESELECTION module

We experiment with various learning rates when training SCIBERT, BioMedRoBERTa, RoBERTa-base, and RoBERTa-large. Below we describe the setting for training RoBERTa-large.

For models trained on SCIFACT, we use an initial learning rate of 1e-5 on the transformer base and 1e-3 on the linear layer. For FEVER + SCIFACT, the learning rate is set to 1e-5 for the entire model for pre-training on FEVER and fine-tuning on SCIFACT. We use a batch size of 256 through gradient accumulation and apply cosine learning rate decay over 20 epochs to find the best performing model on the dev set.

For models trained on FEVER, we set the learning rate to 5e-6 for the transformer base and 5e-5 for the linear layer. For models trained on UKP Snopes, we set the learning rate 1e-5 for the transformer base and 1e-4 for the linear layer. We find that these learning rates help the models converge. We only train the model for 3 epochs on FEVER and 5 epochs on UKP Snopes because they are larger datasets and the models converged within early epochs.

A.3 Training the LABELPREDICTION module

We adopt similar settings as we used for the RATIONALESELECTION module and only change the learning rate to 1e-5 for the transformer base and 1e-4 for the linear layer for models trained on SCIFACT, FEVER, and UKP Snopes. When training on claim / cited abstract pairs labeled NOINFO, we use the k sentences in the abstract with greatest similarity to the claim as rationales (§5). k is sampled from $\{0, 1\}$ with uniform probability.

A.4 Additional training details

All models are trained using a single Nvidia P100 GPU on Google Colabortoary Pro platform.[8] For the RATIONALESELECTION module, it takes about 150 minutes to train on SCIFACT for 20 epochs. 120 minutes on UKP Snopes for 5 epochs, and 700 minutes on FEVER for 3 epochs. For the LABELPREDICTION module, it takes about 130 minutes to train on SCIFACT for 20 epochs, 160 minutes on UKP Snopes for 5 epochs, and 640 minutes on FEVER for 3 epochs.

A.5 Hyperparameter search

The learning rate, batch size, and number of epochs are the most important hyperparameters. We perform manual tuning and select the hyperparameters that produce the highest F1 on the development set. For the learning rate, we experiment with 1e-3, 1e-4, 5e-5, 1e-5, and 5e-6. For batch size, we experiment with 64 and 256. The number of epochs are cutoff after the model converges.

B Statistical analysis

We assess the uncertainty in the results reported in the main results (Table 4) using a simple bootstrap approach (Dror et al., 2018; Berg-Kirkpatrick et al., 2012; Efron and Tibshirani, 1993). Given our test set with $n_{\text{test}} = 300$ claims, we generate $n_{\text{boot}} = 10,000$ bootstrap-resampled test sets by resampling (uniformly, with replacement) n_{test} claims from the test set. For each resampled test set, we compute the metrics in Table 4. Table 6 reports the mean and standard deviation of these metrics, computed over the bootstrap samples. Table 7 reports dev set metrics. Our conclusion that training on SCIFACT improves performance is robust to the uncertainties presented in these tables.

[8] https://colab.research.google.com/

			Sentence-level					
			Selection-Only			Selection+Label		
Retrieval	Model	Row	P	R	F1	P	R	F1
Oracle abstract	Oracle rationale	1	$100.0_{0.0}$	$80.5_{3.3}$	$89.2_{2.1}$	$89.6_{2.7}$	$72.2_{3.7}$	$79.9_{3.0}$
	Zero-shot	2	$42.6_{2.2}$	$45.2_{3.2}$	$43.8_{2.0}$	$36.2_{2.5}$	$38.4_{3.0}$	$37.2_{2.3}$
	VeriSci	3	$76.2_{2.9}$	$63.9_{3.6}$	$69.4_{2.6}$	$66.5_{3.4}$	$55.7_{3.7}$	$60.6_{3.1}$
Open	Oracle rationale	4	$100.0_{0.0}$	$56.6_{4.0}$	$72.2_{3.3}$	$87.6_{3.5}$	$49.5_{3.9}$	$63.2_{3.7}$
	Zero-shot	5	$28.7_{2.3}$	$37.6_{3.4}$	$32.5_{2.3}$	$23.8_{2.3}$	$31.1_{3.1}$	$26.9_{2.3}$
	VeriSci	6	$45.0_{3.0}$	$47.4_{3.8}$	$46.1_{3.0}$	$38.5_{3.0}$	$40.6_{3.6}$	$39.5_{3.0}$

(a) Sentence-level results.

			Abstract-level					
			Label-Only			Label+Rationale		
Retrieval	Model	Row	P	R	F1	P	R	F1
Oracle abstract	Oracle rationale	1	$90.1_{2.2}$	$77.5_{2.8}$	$83.3_{2.4}$	$90.1_{2.2}$	$77.5_{2.8}$	$83.3_{2.4}$
	Zero-shot	2	$86.9_{2.9}$	$53.6_{3.4}$	$66.3_{3.1}$	$67.9_{3.9}$	$41.9_{3.2}$	$51.8_{3.4}$
	VeriSci	3	$87.3_{2.6}$	$65.3_{3.2}$	$74.7_{2.8}$	$84.9_{2.8}$	$63.5_{3.2}$	$72.6_{2.9}$
Open	Oracle rationale	4	$88.9_{2.7}$	$54.1_{3.5}$	$67.2_{3.2}$	$88.9_{2.7}$	$54.1_{3.5}$	$67.2_{3.2}$
	Zero-shot	5	$56.0_{3.9}$	$42.3_{3.4}$	$48.2_{3.3}$	$42.3_{4.0}$	$32.0_{3.2}$	$36.4_{3.3}$
	VeriSci	6	$47.5_{3.3}$	$47.3_{3.5}$	$47.4_{3.1}$	$46.6_{3.3}$	$46.4_{3.5}$	$46.4_{3.1}$

(b) Abstract-level results

Table 6: Test set results as in Table 4, reporting mean and standard deviation over 10,000 bootstrap samples. Standard deviations are reported as subscripts. Some means reported here are slightly different from Table 4 due to sampling variability.

			Sentence-level					
			Selection-Only			Selection+Label		
Retrieval	Model	Row	P	R	F1	P	R	F1
Oracle abstract	Oracle rationale	1	$100.0_{0.0}$	$81.9_{3.2}$	$90.0_{1.9}$	$91.4_{2.5}$	$74.9_{3.6}$	$82.3_{2.9}$
	Zero-shot	2	$40.7_{2.1}$	$48.1_{3.4}$	$44.0_{2.1}$	$36.1_{2.5}$	$42.6_{3.4}$	$39.0_{2.5}$
	VeriSci	3	$79.4_{2.7}$	$59.0_{3.6}$	$67.7_{2.8}$	$71.4_{3.5}$	$53.0_{3.6}$	$60.8_{3.3}$
Open	Oracle rationale	4	$100.0_{0.0}$	$58.4_{4.3}$	$73.7_{3.4}$	$90.2_{3.3}$	$52.7_{4.3}$	$66.4_{3.9}$
	Zero-shot	5	$28.6_{2.0}$	$38.5_{3.6}$	$32.8_{2.3}$	$24.8_{2.2}$	$33.4_{3.4}$	$28.4_{2.4}$
	VeriSci	6	$52.5_{3.5}$	$43.8_{3.7}$	$47.7_{3.2}$	$46.9_{3.7}$	$39.2_{3.6}$	$42.6_{3.2}$

(a) Sentence-level results.

			Abstract-level					
			Label-Only			Label+Rationale		
Retrieval	Model	Row	P	R	F1	P	R	F1
Oracle abstract	Oracle rationale	1	$91.4_{2.2}$	$76.1_{3.0}$	$83.0_{2.5}$	$91.4_{2.2}$	$76.1_{3.0}$	$83.0_{2.5}$
	Zero-shot	2	$88.9_{2.8}$	$58.3_{3.7}$	$70.4_{3.2}$	$69.2_{3.9}$	$45.4_{3.5}$	$54.8_{3.5}$
	VeriSci	3	$91.0_{2.3}$	$67.4_{3.3}$	$77.4_{2.7}$	$85.2_{2.9}$	$63.2_{3.5}$	$72.5_{3.1}$
Open	Oracle rationale	4	$91.0_{2.6}$	$53.1_{3.8}$	$67.0_{3.4}$	$91.0_{2.6}$	$53.1_{3.8}$	$67.0_{3.4}$
	Zero-shot	5	$52.7_{3.7}$	$41.6_{3.7}$	$46.5_{3.4}$	$43.6_{3.7}$	$34.4_{3.5}$	$38.4_{3.3}$
	VeriSci	6	$55.4_{3.7}$	$47.5_{3.6}$	$51.0_{3.3}$	$52.6_{3.7}$	$45.1_{3.6}$	$48.5_{3.3}$

(b) Abstract-level results

Table 7: Dev set results as in Table 4, reporting mean and standard deviation over 10,000 bootstrap samples.

Journal	Count
BMJ	60
Blood	8
Cancer Cell	8
Cell	51
Cell Metabolism	10
Cell Stem Cell	41
Circulation	12
Immunity	33
JAMA	79
Molecular Cell	27
Molecular Systems Biology	5
Nature	29
Nature Cell Biology	26
Nature Communications	19
Nature Genetics	8
Nature Medicine	89
Nature Methods	1
Nucleic Acids Research	10
Plos Biology	36
Plos Medicine	38
Science	7
Science Translational Medicine	2
The Lancet	22
Other	120
Total	741

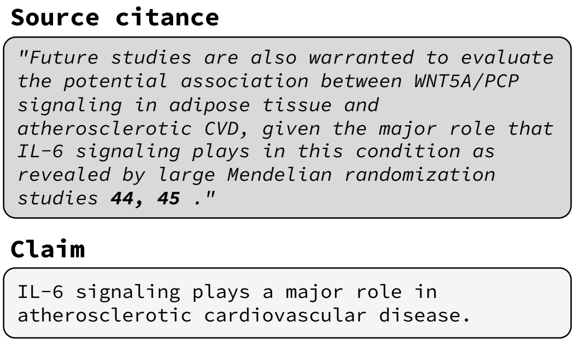
Table 8: Number of cited documents by journal. Some co-cited articles (§3.1) come from journals outside our curated set; these are indicated by "Other".

C Dataset collection and corpus statistics

C.1 Corpus

Source journals Table 8 shows the number of cited abstracts from each of our selected journals. The "Other" category includes "co-cited" (§3.1) abstracts that came from journals not among our pre-defined set.

Distractor abstracts In §3.1, we mention how we increase the size of the corpus by adding *distractor abstracts*. The reason why we do not use the entirety of a large research corpus like S2ORC as our fact-checking corpus is that doing so would introduce many *false negative retrievals*: abstracts containing evidence relevant to a given claim, but not mentioned in the claim's source citance. This can occur either because the citance authors simply were not aware of these abstracts, or because the abstracts were published after the citance was written. These retrievals would be incorrectly marked wrong by our evaluation metrics.



Source citance

*"Future studies are also warranted to evaluate the potential association between WNT5A/PCP signaling in adipose tissue and atherosclerotic CVD, given the major role that IL-6 signaling plays in this condition as revealed by large Mendelian randomization studies **44, 45** ."*

Claim

IL-6 signaling plays a major role in atherosclerotic cardiovascular disease.

Figure 4: A claim written based on a citance. Material unrelated to the citation is removed. The acronym "CVD" is expanded to "cardiovascular disease".

Distractor abstracts as defined in §3.1 have two qualities that make them a good addition to the SciFact corpus: (1) They are cited in the same articles as our evidence abstracts, meaning that they often discuss similar topics and increase the difficulty of abstract retrieval methods based on lexical similarity. (2) The authors of our citances were aware of the distractor abstracts, and chose not to mention them in the citances used to generate claims. This makes them unlikely to be a source of false negative retrievals.

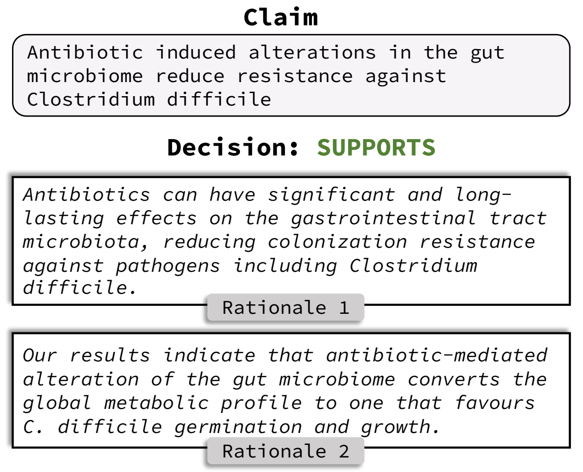
C.2 Annotation examples

Converting citances to claims Figure 4 shows an example of a citance re-written as a claim. The citance discusses the relationship between "atherosclerotic CVD" and "IL-6", and cites two papers (**44** and **45**) as evidence. To convert to a claim, the acronym "CVD" is expanded to "cardiovascular disease", irrelevant information is removed, and the claim is written as an atomic factual statement.

Multiple rationales Figure 5 shows a claim supported by two rationales from the same abstract. The text of each rationale on its own is sufficient to entail the claim.

C.3 Annotators and quality control

Claim writing Student claim writers attended an in-person training session where they were introduced to the task and received in-person feedback from the four experts. Following training, student annotators continued writing claims remotely. The expert annotators monitored claims for quality during the remote annotation process, and provided



Claim

Antibiotic induced alterations in the gut microbiome reduce resistance against Clostridium difficile

Decision: SUPPORTS

Antibiotics can have significant and long-lasting effects on the gastrointestinal tract microbiota, reducing colonization resistance against pathogens including Clostridium difficile.

Rationale 1

Our results indicate that antibiotic-mediated alteration of the gut microbiome converts the global metabolic profile to one that favours C. difficile germination and growth.

Rationale 2

Figure 5: A claim supported by two rationales from the same abstract. The text of each rationale on its own provides sufficient evidence to verify the claim.

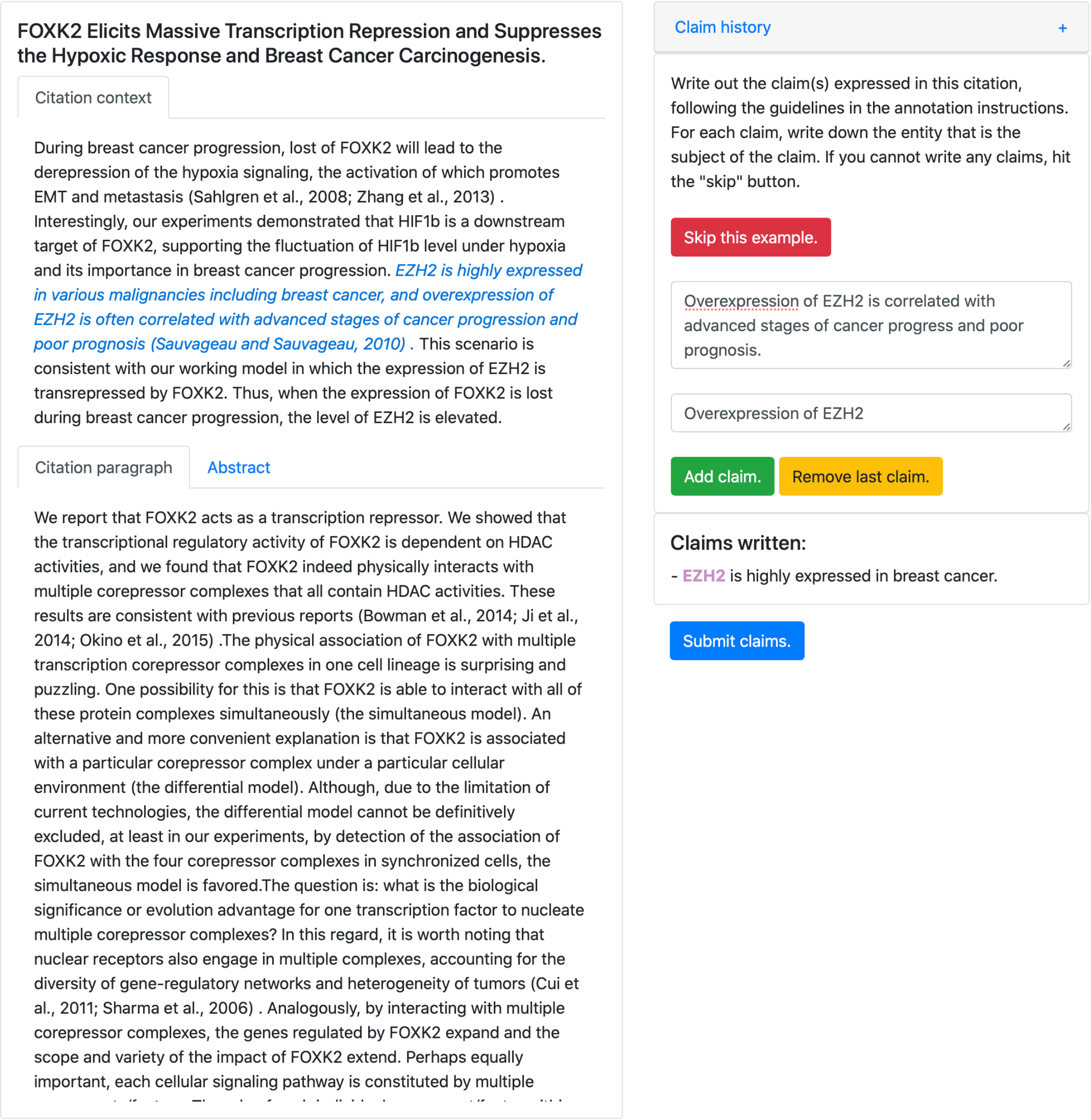
feedback when necessary; low-quality claims were returned to the annotators for re-writing. As a final check, all submitted claims were proofread (and edited if necessary) by an undergraduate whose claims were deemed especially high-quality by the expert annotators.

Claim negations As mentioned in §3.2, an expert annotator wrote claim negations to introduce cases where an abstract REFUTES a claim. The annotator skipped claims that could only be negated by adding obvious triggers like “not”. The majority of claim negations involved a reversal of effect direction; for instance “*A high microerythrocyte count protects against severe anemia*” can be negated as “*A high microerythrocyte count raises vulnerability to severe anemia*”.

Claim verification Annotations were performed remotely through a web interface. Annotators were required to pass a 10-question “quiz” before annotating their own claims. After passing the quiz, subsequent submissions were reviewed by an NLP expert until that expert deemed the annotator reliable. Approved annotators were then assigned to review each others’ submissions. In general, graduate students were assigned to review annotations from undergraduates.

D Annotation interfaces and guidelines

We show a screenshot of the claim writing interface in Figure 6, and the claim verification interface in Figure 7. The complete annotation guide for claim verification is available at the following URL: https://scifact.s3-us-west-2.amazonaws.com/doc/evidence-annotation-instructions.pdf.



FOXK2 Elicits Massive Transcription Repression and Suppresses the Hypoxic Response and Breast Cancer Carcinogenesis.

During breast cancer progression, lost of FOXK2 will lead to the derepression of the hypoxia signaling, the activation of which promotes EMT and metastasis (Sahlgren et al., 2008; Zhang et al., 2013) . Interestingly, our experiments demonstrated that HIF1b is a downstream target of FOXK2, supporting the fluctuation of HIF1b level under hypoxia and its importance in breast cancer progression. *EZH2 is highly expressed in various malignancies including breast cancer, and overexpression of EZH2 is often correlated with advanced stages of cancer progression and poor prognosis (Sauvageau and Sauvageau, 2010) .* This scenario is consistent with our working model in which the expression of EZH2 is transrepressed by FOXK2. Thus, when the expression of FOXK2 is lost during breast cancer progression, the level of EZH2 is elevated.

We report that FOXK2 acts as a transcription repressor. We showed that the transcriptional regulatory activity of FOXK2 is dependent on HDAC activities, and we found that FOXK2 indeed physically interacts with multiple corepressor complexes that all contain HDAC activities. These results are consistent with previous reports (Bowman et al., 2014; Ji et al., 2014; Okino et al., 2015) .The physical association of FOXK2 with multiple transcription corepressor complexes in one cell lineage is surprising and puzzling. One possibility for this is that FOXK2 is able to interact with all of these protein complexes simultaneously (the simultaneous model). An alternative and more convenient explanation is that FOXK2 is associated with a particular corepressor complex under a particular cellular environment (the differential model). Although, due to the limitation of current technologies, the differential model cannot be definitively excluded, at least in our experiments, by detection of the association of FOXK2 with the four corepressor complexes in synchronized cells, the simultaneous model is favored.The question is: what is the biological significance or evolution advantage for one transcription factor to nucleate multiple corepressor complexes? In this regard, it is worth noting that nuclear receptors also engage in multiple complexes, accounting for the diversity of gene-regulatory networks and heterogeneity of tumors (Cui et al., 2011; Sharma et al., 2006) . Analogously, by interacting with multiple corepressor complexes, the genes regulated by FOXK2 expand and the scope and variety of the impact of FOXK2 extend. Perhaps equally important, each cellular signaling pathway is constituted by multiple

Write out the claim(s) expressed in this citation, following the guidelines in the annotation instructions. For each claim, write down the entity that is the subject of the claim. If you cannot write any claims, hit the "skip" button.

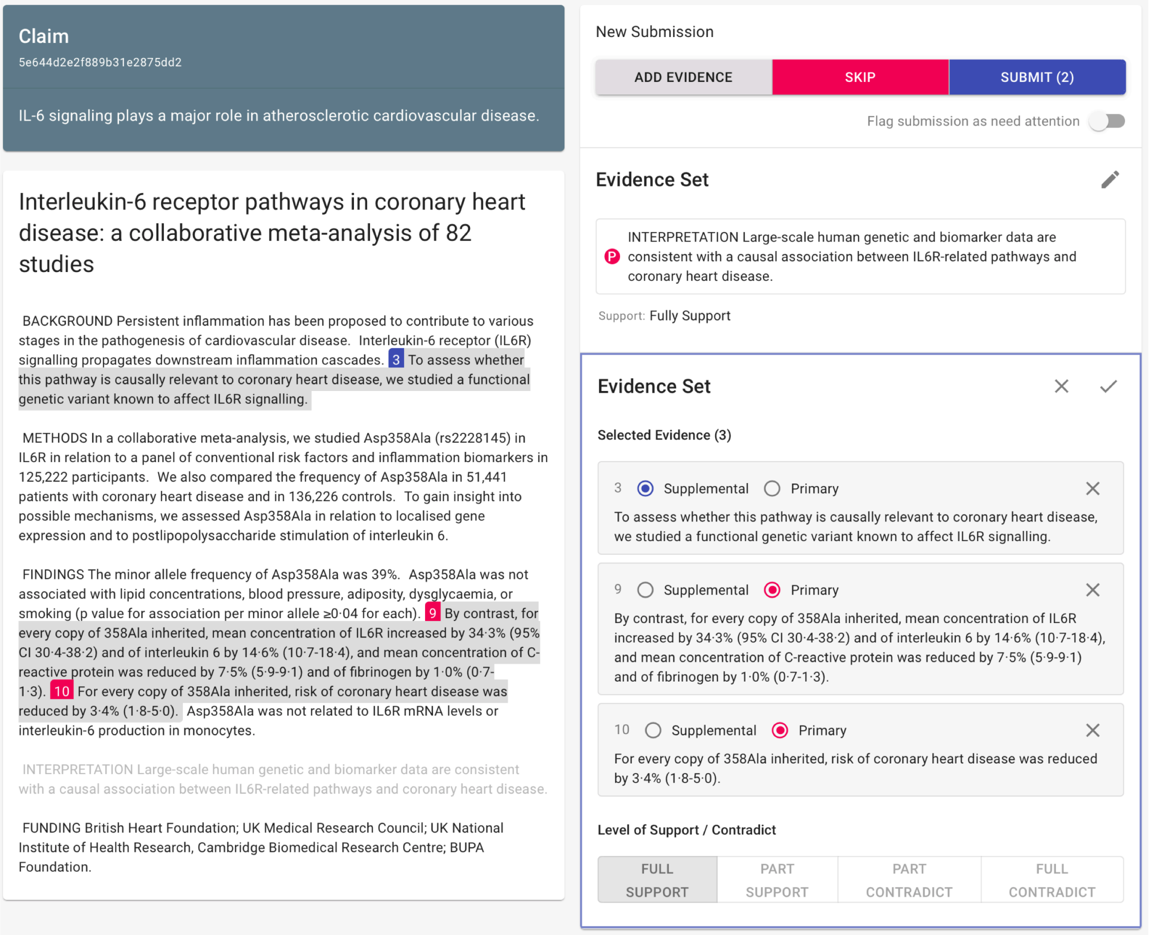
Overexpression of EZH2 is correlated with advanced stages of cancer progress and poor prognosis.

Overexpression of EZH2

Claims written:

- EZH2 is highly expressed in breast cancer.

Figure 6: The claim-writing interface. The citation sentence is highlighted in blue on the top left. Additional context is provided on bottom left. The right side shows two claims that could be written based on this citation sentence.



IL-6 signaling plays a major role in atherosclerotic cardiovascular disease.

Interleukin-6 receptor pathways in coronary heart disease: a collaborative meta-analysis of 82 studies

BACKGROUND Persistent inflammation has been proposed to contribute to various stages in the pathogenesis of cardiovascular disease. Interleukin-6 receptor (IL6R) signalling propagates downstream inflammation cascades. 3 To assess whether this pathway is causally relevant to coronary heart disease, we studied a functional genetic variant known to affect IL6R signalling.

METHODS In a collaborative meta-analysis, we studied Asp358Ala (rs2228145) in IL6R in relation to a panel of conventional risk factors and inflammation biomarkers in 125,222 participants. We also compared the frequency of Asp358Ala in 51,441 patients with coronary heart disease and in 136,226 controls. To gain insight into possible mechanisms, we assessed Asp358Ala in relation to localised gene expression and to postlipopolysaccharide stimulation of interleukin 6.

FINDINGS The minor allele frequency of Asp358Ala was 39%. Asp358Ala was not associated with lipid concentrations, blood pressure, adiposity, dysglycaemia, or smoking (p value for association per minor allele ≥0·04 for each). 9 By contrast, for every copy of 358Ala inherited, mean concentration of IL6R increased by 34·3% (95% CI 30·4-38·2) and of interleukin 6 by 14·6% (10·7-18·4), and mean concentration of C-reactive protein was reduced by 7·5% (5·9-9·1) and of fibrinogen by 1·0% (0·7-1·3). 10 For every copy of 358Ala inherited, risk of coronary heart disease was reduced by 3·4% (1·8-5·0). Asp358Ala was not related to IL6R mRNA levels or interleukin-6 production in monocytes.

INTERPRETATION Large-scale human genetic and biomarker data are consistent with a causal association between IL6R-related pathways and coronary heart disease.

FUNDING British Heart Foundation; UK Medical Research Council; UK National Institute of Health Research, Cambridge Biomedical Research Centre; BUPA Foundation.

P INTERPRETATION Large-scale human genetic and biomarker data are consistent with a causal association between IL6R-related pathways and coronary heart disease.

Support: Fully Support

3 ● Supplemental ○ Primary

To assess whether this pathway is causally relevant to coronary heart disease, we studied a functional genetic variant known to affect IL6R signalling.

9 ○ Supplemental ● Primary

By contrast, for every copy of 358Ala inherited, mean concentration of IL6R increased by 34·3% (95% CI 30·4-38·2) and of interleukin 6 by 14·6% (10·7-18·4), and mean concentration of C-reactive protein was reduced by 7·5% (5·9-9·1) and of fibrinogen by 1·0% (0·7-1·3).

10 ○ Supplemental ● Primary

For every copy of 358Ala inherited, risk of coronary heart disease was reduced by 3·4% (1·8-5·0).

Figure 7: The evidence collection interface.