PubHealthTab: A Public Health Table-based Dataset for Evidence-based Fact Checking

Mubashara Akhtar, Oana Cocarascu and **Elena Simperl**
Department of Informatics, King's College London
{mubashara.akhtar,oana.cocarascu,elena.simperl}@kcl.ac.uk

Abstract

Inspired by human fact checkers, who use different types of evidence (e.g. tables, images, audio) in addition to text, several datasets with tabular evidence data have been released in recent years. Whilst the datasets encourage research on table fact-checking, they rely on information from restricted data sources, such as Wikipedia for creating claims and extracting evidence data, making the fact-checking process different from the real-world process used by fact checkers. In this paper, we introduce *PubHealthTab*, a table fact-checking dataset based on real-world public health claims and noisy evidence tables from sources similar to those used by real fact checkers. We outline our approach for collecting evidence data from various websites and present an in-depth analysis of our dataset. Finally, we evaluate state-of-the-art table representation and pre-trained models fine-tuned on our dataset, achieving an overall F_1 score of 0.73.

1 Introduction

Fact-checking is the task of establishing the veracity of factual information, commonly performed manually by journalists. In addition to classifying how truthful claims are, human fact checkers also provide evidence for their judgements. To support this process with computational tools, researchers have compiled several datasets for evidence-based automated fact-checking (AFC), which include information about the sources supporting or refuting the claims alongside veracity labels (Thorne et al., 2018; Chen et al., 2020b; Aly et al., 2021; Schuster et al., 2021; Nørregaard and Derczynski, 2021).

While a large share of the datasets used in evidence-based AFC focus on textual evidence (e.g. (Thorne et al., 2018; Augenstein et al., 2019; Diggelmann et al., 2020; Schuster et al., 2021)), some recent datasets also cover structured data, for instance in the form of web tables (Chen et al., 2020b; Aly et al., 2021). This is useful, as human fact checkers often need to consider a range of data modalities to verify claims. However, two main limitations remain. First, existing table fact-checking datasets consist largely of claims which have been 'artificially' created via online crowdsourcing, starting from randomly selected evidence tables. Second, the datasets use single sources of evidence, for instance Wikipedia; this is different from how human fact checkers go about the task - more often than not, they consult multiple primary sources, including websites, databases, and public reports.

To overcome these limitations, we propose *PubHealthTab*[1], a new table fact-checking dataset, using the *PubHealth* dataset (Kotonya and Toni, 2020) as a seed. PubHealth has a number of advantages. It contains public health claims that human fact-checkers work on. The authors compared the complexity of these claims to real-world political claims, as well as to claims created by crowdworkers (Kotonya and Toni, 2020). As a proxy for complexity, they determined the reading skills needed to understand the claims. They established that public health claims are much more challenging, requiring high school levels of reading of 10 to 12 rather than 6 to 8 for political and crowdsourced claims. PubHealth also includes multiple sources of evidence for the claims, however, the evidence is purely text-based. In our dataset, we include web tables as evidence, extracted from different websites, similar to those used by human fact-checkers.

We designed a hybrid dataset pipeline, which takes PubHealth claims and links them, via Wikipedia articles, to other websites containing potential evidence tables. We used crowdsourcing in three ways: to establish the relevance of the extracted tables; to adjust PubHealth claims to support or refute the tables; and finally to assess the

[1] https://github.com/mubasharaak/PubHealthTab

Findings of the Association for Computational Linguistics: NAACL 2022, pages 1 - 16
July 10-15, 2022 ©2022 Association for Computational Linguistics

quality of the new claims. The result is a dataset of $1,942$ claim-table pairs about public health, drawing on evidence from more than 300 websites.

We analysed the dataset to spot potential biases in the way we collected the data and compared PubHealthTab with other table-based fact-checking datasets. Moreover, we experimented with several BERT-based models and table representations to understand how our dataset performs on state-of-the-art AFC, achieving an overall F_1 score of 0.73. Both allowed us to identify areas of future improvement, in particular to refute claims against evidence consisting of mostly numerical data or with noisy text headers.

2 Background & Related Work

2.1 Evidence-based Fact-Checking

Evidence-based AFC requires one to predict a veracity label against the evidence. While most datasets focus on textual sources of evidence (Thorne et al., 2018; Jiang et al., 2020; Diggelmann et al., 2020; Schuster et al., 2021), human fact checkers use a wider range of modalities (Nakov et al., 2021). To verify factual information, they commonly ask experts, search in databases, and consult text, tables, and graphics from a multitude of sources, including scholarly literature, public reports, and official statistics.[2]

2.2 Table Fact-Checking Datasets

There is a small number of datasets that consider tables in AFC. However, in all cases, the claims are created by crowdworkers given evidence from Wikipedia. For instance, TabFact (Chen et al., 2020b) contains tables extracted from Wikipedia and considers two classes for the claim veracity: entailment and contradiction. The InfoTabs dataset (Gupta et al., 2020) has claims that can be verified using information from Wikipedia info-boxes, with an additional "neutral" class. In FEVEROUS (Aly et al., 2021), claims are verified using text, tables, and lists from Wikipedia. Finally, the recent Sem-Eval fact-checking challenge, Sem-Tab-Facts (Wang et al., 2021), released a table fact-checking dataset with tables extracted from scientific articles. Claims were created by crowd workers based on sentences in the article describing these tables.

[2] https://ballotpedia.org/The_methodologies_of_fact-checking

2.3 Tables in Other NLP Tasks

There is an increasing body of literature looking at tables alongside text for NLP tasks such as table question answering (tableQA) or table-to-text natural language generation (NLG). The former aims to find answers to natural language questions in tabular data (Pasupat and Liang, 2015; Zhong et al., 2017; Iyyer et al., 2017) and inspired the first table fact-checking dataset (Chen et al., 2020b). Researchers later introduced variations of the task with additional modalities (Chen et al., 2020c; Hannan et al., 2020) or sub-tasks such as table retrieval (Chen et al., 2021). There are also several table-to-text NLG datasets, for instance numericNLG (Suadaa et al., 2021) with tables extracted from scientific papers, and LogicNLG (Chen et al., 2020a) with Wikipedia tables. We used some of the methods proposed by the numericNLG team (Suadaa et al., 2021) to represent tables in our experiments.

2.4 The PubHealth Dataset

As noted earlier, we used PubHealth (Kotonya and Toni, 2020) as a starting point for creating our table fact-checking dataset. PubHealth consists of real-world claims about public health extracted from fact-checking and news review websites. The authors comment that the majority of fact-checking datasets either concentrate on politics (Wang, 2017; Augenstein et al., 2019) or are built for research purposes (Thorne et al., 2018; Chen et al., 2020b). Each record in the PubHealth dataset consists of a claim, the full text of the fact-checking or news article, which discusses its veracity, and the article summary or a justification for the veracity label.

3 The PubHealthTab Dataset

Figure 1 shows an overview of the data construction pipeline. In the top half, we automatically create pairs of claims and tables. We start from the PubHealth claims, assess them for relevance and then match the remaining ones with web tables (see Section 3.1). In the bottom half, we use crowdsourcing to filter tables, adjust claims to tables, and check for quality (see Section 3.1.2).

3.1 Dataset Construction

3.1.1 Steps 1 to 3: From Claims to Tables

In Step 1 we removed ambiguous and out-of-domain claims from the PubHealth dataset using a

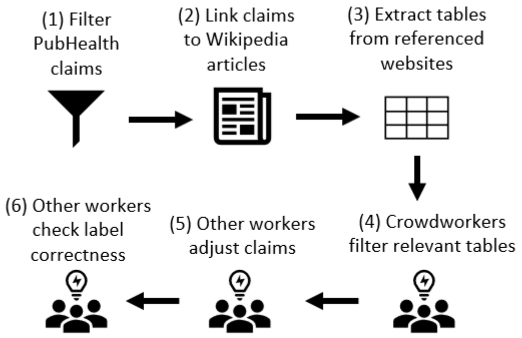


Figure 1: Dataset creation process.

lexicon of 4132 medical terms from: Wikipedia;[3] medical dictionaries from Harvard University,[4] University of Michigan[5], and Schulich School of Medicine and Dentistry[6]; as well as the ConceptNet knowledge graph.[7] We retained the claims that contained at least one token matching the lexicons. For the other claims, we carried out NER to detect medical entities that the lexicons might have missed, using SciSpacy (Neumann et al., 2019). We kept the claims for which we could find an entity in the claim text whose ConceptNet node was liked to a lexicon term via the "hasContext" relation.[8]

In Step 2 we linked the claims from Step 1 to Wikipedia articles using two entity linking services: ELQ (Li et al., 2020) and WAT,[9] for better coverage. We then took the websites referenced by the articles as a source of evidence tables. In Step 3, from all Wikipedia references, we kept those in English that could be scraped and which contained at least one table HTML tag ($\langle table \rangle$). We heuristically removed all tables that were used purely for formatting reasons, and then ranked the remaining tables based on their BM25 similarity to the claim text. The result of this step was a set of 1915 claim-table pairs (1010 claims and 1422 tables from 1196 websites), which was fed to the crowdsourced half of the pipeline.

3.1.2 Steps 4 to 6: Crowdsourcing

We ran three crowdsourcing tasks on Amazon Mechanical Turk (MTurk) in May-June 2021: *table relevance*, *claim adjustment*, and *verification*, loosely following the *"find-fix-verify"* crowdsourcing workflow for text processing by Bernstein et al. (2015). For each of the three tasks, we checked for quality, evaluated worker agreement, and aggregated the results before feeding them to the subsequent task.

Recruitment and training of workers. We allocated each task to three crowdworkers. Only workers with minimum 1000 previously-approved tasks and an approval rate of 95% or above were eligible to work on the tasks. Moreover, all workers had to pass a table literacy qualification test (see appendix). To train the workers, we followed the recommendations from Gadiraju et al. (2015); Doroudi et al. (2016) and included examples of expert-labelled tasks in the instructions, including the rationales for the chosen labels.

Tasks design. The tasks were designed as follows (see appendix for instructions and interfaces):

1. Task 1 - table relevance: We asked crowdworkers if claims and tables were related to each other. This was needed to evaluate the ranked list of tables from Step 3 (Figure 1), where we matched claims to tables using BM25. For each claim-table pair, workers could choose between four options: table *supports*, *refutes*, *is related but more information is needed*, and *is unrelated* to the claim. In addition, we also asked the crowd to name the columns which contributed to their choice. Each task had seven claim-table pairs, of which two were from the gold standard (see quality assurance below). We used majority voting to aggregate the answers.

2. Task 2 - claim adjustment: The input for this task were only the claim-table pairs which were judged as *related but not enough information* in the previous step. We asked crowdworkers to adjust a claim so that they could be supported or refuted by the table. The workers also had to flag whether the table supported or refuted the claim. Each task consisted of five claim-table pairs. As this was an open-

[3] https://en.wikipedia.org/wiki/Glossary_of_medicine
[4] https://www.health.harvard.edu/a-through-c
[5] https://apps.lib.umich.edu/medical-dictionary/
[6] https://www.schulich.uwo.ca/pathol/about_us/resources/glossary_of_medical_terms.html
[7] https://conceptnet.io/
[8] https://github.com/commonsense/conceptnet5/wiki/Relations
[9] https://sobigdata.d4science.org/web/tagme/wat-api

	K-α	F-κ	R-κ
Table relevance	0.26	0.38	0.65
Verification	0.60	0.60	0.67

Table 1: Inter-annotator agreement scores for the *table relevance* task and the *verification* task.

ended task, we evaluated the results in the third crowdsourcing task.

3. Task 3 - verification: We asked crowdworkers to verify the adjusted claims. Again, each task had seven pairs of claims and tables, with two gold pairs. Workers could choose between four labels: *supports*, *refutes*, *related but not enough information*, and *unrelated*. We performed majority voting to aggregate the answers.

For the final dataset (see Section 3.2), we discarded the pairs of adjusted claims and tables labelled as *unrelated* by the majority of workers.

Quality assurance. For each task, we followed best practices to maintain annotation quality and detect malicious behaviour. One of the authors created a gold standard of 30 claim-table pairs for the close-ended tasks (table relevance and verification); we used two gold pairs per task. Workers who failed those two gold pairs could not submit their work. For the remaining submissions, we computed the *inter-annotator agreement*.

Table 1 shows the inter-annotator agreement scores using Krippendorff's alpha (K-α), Fleiss' kappa (F-κ), and Randolph's kappa (R-κ). F-κ is prone to the high agreement but low kappa phenomenon when the dataset is imbalanced (Feinstein and Cicchetti, 1990); this was the case for the table relevance task: after aggregating the answers with majority voting, we had the following distribution: less than 1% *support*, less than 1% *refute*, 22% *related but not enough information*, and 77% *unrelated*. This is why we used R-κ, which yields more accurate results for imbalanced data. For the verification task, the data was more balanced, which is reflected in the similar scores. For both tasks, we obtained a R-κ value of at least 0.65, which indicates substantial agreement according to Landis and Koch (1977).

The claim adjustment task was open-ended. We allowed only submissions which met a set of criteria, for instance by looking at the time spent per task and comparing the original and adjusted claim; the full list of criteria is in the appendix. We also manually inspected the adjusted claims before accepting them. We randomly sampled one claim for each submission and accepted the work if its quality was sufficient. After a first pilot round, we banned workers with malicious behaviour, e.g. workers who did not adjust the claims, but only added or removed one token.

Claim: Measles outbreak in Quebec carries a different strain than that in Ontario in recent years.

Province	Year	Number of cases	Duration (weeks)	Strain
Quebec	2007	94	24	D4
Ontario	2008	53	11	D8
British Columbia	2010	82	7	D8 and H1
Quebec	2011	20	11	D4
Quebec	2011	678	33	D4

Caption: Table 1. Measles Outbreaks in Canada, by province, 2007 to 2011
Website title: Guidelines for measles outbreak in Canada - Canada.ca
Veracity label: SUPPORTS
Source: *https://www.canada.ca/en/public-health/services/reports-publications/canada-communicable-disease-report-ccdr/monthly-issue/2013-39/guidelines-prevention-control-measles-outbreaks-canada.html*

Figure 2: A *support* example from PubHealthTab.

3.2 Dataset Statistics

Our PubHealthTab dataset comprises $1,942$ claim-table pairs. A claim is a natural language sentence checked against a table. Each pair is labelled as *support*, *refute*, or *not enough information (NEI)*, following Thorne et al. (2018); Gupta et al. (2020); Diggelmann et al. (2020); Aly et al. (2021). The dataset has $1,019$ supported claims, 462 refuted claims, and 461 NEI claims. Figure 2 shows an example.

The evidence table is organised as a list of n rows. Each row is a list of cells, where m, the number of cells, can vary across rows. If the first row is a header, it is instead saved as "header_horizontal". Similarly, if the first column is a header, it is saved as "header_vertical". For each table, we provide the source website and, if available, the table caption. Moreover, each record also includes the original PubHealth claim text, which was adjusted by crowdworkers in Step 5 (Figure 1).

Table 2 compares the original PubHealth dataset with our dataset, PubHealthTab.

	PubHealthTab	PubHealth
Entries	1,942	11,832
Evidence type	Table	Text
Claim length	20 - 194	25 - 400
Veracity labels	{supports, refutes, NEI}	{true, mixture, false, unproven}

Table 2: Comparison between our dataset and PubHealth (Kotonya and Toni, 2020).

	Adj. claim length	Veracity label
Table length	0.05 (Pearson)	0.35 (F-test)
Adj. claim length	-	0.47 (F-test)
Caption available	0.36 (F-test)	0.05 (χ^2 test)
Header available	0.16 (F-test)	0.03 (χ^2 test)

Table 3: Calculated p-values for the significance tests.

	Bigram b	LMI	$p(l,b)$	*count*
Supported claims	the highest	1009	0.86	44
	has the	989	0.8	60
	percentage of	579	0.88	24
	had a	423	0.88	17
	highest number	418	0.93	14
	there is	376	0.79	24
	more than	364	0.73	37
Refuted claims	found on	1030	0.61	28
	breast cancer	617	0.46	35
	is found	599	0.48	29
	be found	493	0.62	13
	on page	471	0.42	36
	is about	450	0.64	11
	has a	433	0.34	86
NEI claims	the table	675	0.46	13
	of domestic	621	0.8	5
	health care	584	0.25	36
	domestic violence	564	0.67	6
	in a	516	0.57	7
	for health	398	0.6	5
	to the	365	0.28	18

Table 4: Top LMI-ranked bigrams for support, refute and NEI claims (including probability and count).

4 Dataset Analysis

We analysed the PubHealthTab dataset for biases and correlations, and compared it to other table fact-checking datasets. We applied three methods: (i) correlation analysis of table attributes; (ii) Local Mutual Information (LMI) on adjusted claims; and (iii) claim-only veracity prediction.

4.1 Correlation analysis of table attributes

While correlations between claims and veracity labels in fact-checking datasets have been previously explored (Schuster et al., 2019; Aly et al., 2021; Thorne et al., 2021), such underlying relationships might also be present in the evidence data. Thus, we examined correlations related to tables in the PubHealthTab dataset. We analysed if the veracity labels and the length of adjusted claims were correlated with the following table attributes that were visible to crowdworkers during annotation: table length (i.e. number of rows), availability of table captions, and availability of table headers.

Depending on the type of the attribute analysed, we used: the Pearson correlation coefficient, the χ^2 test, and the Anova F-test and a significance level α of 0.05 to examine correlations. The p-values for all attribute pairs are shown in Table 3. No significant correlations were found between the adjusted claim length and the table attributes' length, caption availability, and header availability. Given p-values $\geq \alpha$, the hypothesis of independence holds for these pairs of variables. Similarly, the veracity labels were not significantly correlated with the table length, caption availability, and adjusted claim length. For the correlation between veracity labels and header availability, we calculated a p-value of 0.03 indicating an underlying relationship between the variables. Examining the attributes in detail, we found that tables with headers were more prominent for supported and refuted claims than for NEI claims in the PubHealthTab dataset.

4.2 Local Mutual Information

Following Schuster et al. (2019), we analysed the correlation between frequently occurring phrases in adjusted claims and their veracity labels. We computed the Local Mutual Information (LMI) score (Evert, 2005) between a bigram b and the claim's veracity label l: $LMI(b,l) = p(b,l) * log(\frac{p(l|b)}{p(l)})$. Unlike the Point-wise Mutual Information (PMI) score, $PMI = log(\frac{p(l|b)}{p(l)})$, the LMI score avoids over-weighting bigrams with no or low occurrences in the overall dataset by multiplying it with the probability $p(b,l)$, where $p(b,l)$ is approximated by $\frac{count(b,l)}{|B|}$, $|B|$ is the number of all bigrams in the dataset and $count(b,l)$ is the number of times b and l occur together.

Table 4 shows the top LMI-ranked bigrams for PubHealthTab claims. We found similar bigrams in different classes, for example "has a" appears in refuted claims and "had a" in supported claims. Furthermore, no top-ranked bigram of refuted claims contains negation tokens such as "not", "never" or "false". Thus, we conclude that the top-ranked bigrams occurring in claims are not specific to their veracity labels.

4.3 Claim-only Veracity Prediction

We fine-tuned a BERT base model (Devlin et al., 2019) on PubHealthTab claims to predict their veracity labels using only the text as input and ignoring evidence tables. A claim-only model that performs well could indicate underlying correlations between the claims and the veracity labels. A similar approach was used by Schuster et al. (2019) to evaluate claim-only biases in the FEVER dataset (Thorne et al., 2018). Using the fine-tuned claim-only BERT model, we obtain an F_1 score of 0.51 on our test set. Comparing the F_1 score of the claim-only model to the performance of models using evidence data (see Section 5), we conclude that claims alone are not sufficient for the BERT model to predict the veracity labels.

4.4 Table Analysis

We compared PubHealthTab to three fact-checking datasets that use tables, TabFact, InfoTabs, and FEVEROUS (Table 5). Whilst almost all TabFact, InfoTabs and FEVEROUS tables have headers, this is not the case in more than half (56.9%) of PubHealthTab tables. Similarly, all TabFact and InfoTabs tables include captions and approximately only one-fifth of PubHealthTab tables (21%) and FEVEROUS tables (22%) have captions. While captions and headers can be useful for understanding the context of a table, these attributes are not always present in real-world tables.

The average number of characters per cell is 13.4 for PubHealthTab tables, more than the average cell length of TabFact tables (8.6) and less than for InfoTabs (22.6) and FEVEROUS (17.3). Moreover, PubHealthTab tables show the highest ratio of cells with numerical content (59%) and the smallest ratio with text-only content compared to the other datasets. Numerical content can pose a challenge for state-of-the-art NLP models as previous works have shown (Suadaa et al., 2021).

5 Experiments and Results

We experimented with several table representation techniques and state-of-the-art models on PubHealthTab to understand related challenges.

5.1 Table Representation

To assess the impact of different table representation methods on the table fact-checking task, we used five table representation techniques. We also used the BERT-based TAPAS model which extends the BERT model architecture with three additional embeddings to encode table structure. We describe the TAPAS model in more detail when we discuss the modelling approaches in Section 5.2. We describe the table representations in detail below:

Concatenation: transforms the entire content of a table into one flat string ignoring the table structure. The table caption, headers, and content are concatenated and used jointly as input for label prediction.

Template-based concatenation: maps table columns and cell values into a structured form using the following template applied to each row: `row_1: column_1:cell_value, column_2:cell_value, [...]`. The `row` and `column` tokens were replaced by the corresponding vertical header (for row) and horizontal header (for column), if available.

Template-based sentences: We defined a template to convert table content to one sentence per row. For example, given a table with headers "medicine" and "price", and two cells in the first row, we generate the following template-based sentence for this row: *In row one column one (medicine) is Panadol, column two (price) is £15.*

T5 (concatenation): Similarly to Suadaa et al. (2021), we used text from representation *concatenation* as input to the T5 text generation model (Raffel et al., 2020) to generate sentences that describe the tables.

T5 (template): We used text from representation *template-based sentence* as input to the T5 model.

5.2 Modelling Approaches

Based on the previously described table representation methods, we evaluated state-of-the-art NLP models on PubHealthTab. We use models previously applied in table fact-checking (BERT, ALBERT, RoBERTa) (Chen et al., 2020b; Gupta et al., 2020; Aly et al., 2021), as well as domain-specific models (BioBERT, BlueBERT, ClinicalBERT), pretrained on large-scale health datasets. We describe the models below:

BERT: We used the uncased BERT-base (Devlin et al., 2019) model from `huggingface` library[10].

ALBERT: A transformer-based model that extends BERT with a parameter-reduction technique, resulting in lower memory consumption and higher training speed (Lan et al., 2020).

[10] `https://huggingface.co`

	Our Dataset	TabFact	InfoTabs	FEVEROUS
Total number of tables	1,942	16,573	2,540	28,760
% of tables with caption	21%	100%	100%	22%
% of tables with header	56.9%	100%	100%	97%
% of tables with <5 rows	23.1%	0.1%	7.5%	18%
% of tables with =>5 rows & <= 10 rows	53.8%	40.7%	56%	44%
% of tables with >10 rows	23.1%	59.2%	36.5%	38%
Ratio of cells with only string content	30.6%	40.1%	45.8%	34%
Ratio of cells with numerical content	59%	53.6%	35.5%	40%
Avg number of characters per cell	13.4	8.6	22.6	17.3

Table 5: Comparison of table fact checking datasets.

	Train	Valid	Test	Sum
Support	810	106	103	1019
Refute	370	46	46	462
NEI	373	43	45	461
Sum	1553	195	194	1942

Table 6: Class distribution across dataset split.

RoBERTa: We used the RoBERTa-Large model released by Nie et al. (2020). The model was pre-trained on SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018), ANLI (Nie et al., 2020), and FEVER (Thorne et al., 2018).

BioBERT: A domain-specific BERT model, pre-trained on PubMed abstracts and PMC full-text articles (Lee et al., 2020). The model was fine-tuned on two NLI datasets, SNLI and MultiNLI.

BlueBERT: The model was pre-trained on PubMed abstracts and MIMIC-III clinical notes, a database of electronic health records from ICU patients at a Boston hospital (Peng et al., 2019).

ClinicalBERT: A BERT model which was pre-trained on MIMIC-III data (Huang et al., 2019).

TAPAS: An extension to BERT which uses additional, table-specific embeddings (column embeddings, row embeddings, rank embeddings) that capture the table structure (Herzig et al., 2020). We experiment with TAPAS on our dataset as it achieved good performance on the TabFact dataset.

We partitioned the dataset into training (80%), test (10%), and validation (10%) sets. Table 6 shows the class distribution across the dataset split. We performed hyper-parameter search on the validation set and evaluated the following parameters for each model before selecting the best-performing combination: $\{4, 8, 16\}$ for batch size, $\{$1e-3, 1e-5, 1e-7$\}$ for learning rate, $\{2, 3, 4, 5\}$ for training epochs, and $\{0.01, 0.001, 0.0001\}$ for weight decay.

	Represent.	All	Sup.	Ref.	NEI
BERT	concatenation	**0.60**	0.72	0.28	0.81
	template sent.	0.57	0.78	0.04	0.89
	template concat.	0.57	0.75	0.11	0.85
	T5 concat.	0.55	0.75	0.07	0.83
	T5 template	0.53	0.71	0.03	0.84
ALBERT	concatenation	0.55	0.72	0.15	0.79
	template sent.	**0.58**	0.69	0.27	0.79
	template concat.	0.55	0.71	0.17	0.78
	T5 concat.	0.54	0.74	0.07	0.83
	T5 template	0.55	0.75	0.11	0.79
RoBERTa	concatenation	0.69	0.79	0.44	0.84
	template sent.	0.70	0.77	0.48	0.84
	template concat.	0.66	0.75	0.39	0.84
	T5 concat.	**0.73**	0.78	0.52	0.89
	T5 template	0.68	0.74	0.45	0.84
BioBERT	concatenation	0.57	0.68	0.29	0.76
	template sent.	**0.60**	0.71	0.33	0.76
	template concat.	0.58	0.68	0.3	0.75
	T5 concat.	0.58	0.68	0.33	0.73
	T5 template	0.58	0.71	0.30	0.74
BlueBERT	concatenation	0.50	0.72	0.04	0.77
	template sent.	**0.56**	0.71	0.23	0.74
	template concat.	0.54	0.69	0.20	0.75
	T5 concat.	0.52	0.70	0.13	0.72
	T5 template	0.54	0.68	0.22	0.72
ClinicalBERT	concatenation	0.51	0.75	0	0.78
	template sent.	**0.58**	0.72	0.20	0.83
	template concat.	**0.58**	0.74	0.19	0.80
	T5 concat.	0.55	0.76	0.10	0.80
	T5 template	0.55	0.73	0.13	0.78
	TAPAS	0.48	0.67	0.28	0.48

Table 7: F_1 (macro) score for different state-of-the-art models and table representations on PubHealthTab.

5.3 Discussion

We evaluated and compared the table representation and modelling approaches, and report the overall (macro) F_1 score and the F_1 scores for each class in Table 7.

Table Representations. The resulting F_1 scores across all models and veracity classes remained overall the same when different methods for table representation were applied. The *template-based*

	Dataset	All	Sup.	Ref.	NEI
Concat.	PubHealthTab	0.69	0.79	0.44	0.84
	InfoTabs	0.78	0.78	0.76	0.81
	TabFact	0.49	0.34	0.65	-
	FEVEROUS	0.68	0.89	0.87	0.29
T. sent.	PubHealthTab	0.70	0.77	0.48	0.84
	InfoTabs	0.77	0.77	0.73	0.81
	TabFact	0.44	0.23	0.65	-
	FEVEROUS	0.66	0.88	0.85	0.27
T. concat.	PubHealthTab	0.66	0.75	0.39	0.84
	InfoTabs	0.78	0.78	0.75	0.81
	TabFact	0.50	0.36	0.65	-
	FEVEROUS	0.67	0.88	0.86	0.26
T5 concat.	PubHealthTab	0.73	0.78	0.52	0.89
	InfoTabs	0.73	0.72	0.69	0.77
	TabFact	0.47	0.29	0.65	-
	FEVEROUS	0.64	0.86	0.83	0.22
T5 temp.	PubHealthTab	0.68	0.74	0.45	0.84
	InfoTabs	0.72	0.72	0.68	0.77
	TabFact	0.46	0.25	0.67	-
	FEVEROUS	0.64	0.86	0.83	0.24

Table 8: F_1 score for RoBERTa with different representation methods on various table fact-checking datasets.

sentence approach outperforms other representation techniques in terms of the overall F_1 score for four out of six models (i.e. ALBERT, BioBERT, BlueBERT, and ClinicalBERT). However, for all four models, the difference to the second highest scoring representation was relatively small, between 0.02 and 0.03. Thus, choosing between *concatenation* and *template* did not seem to influence the overall claim classification.

Models. RoBERTa outperformed the other models across all representations, followed by BioBERT. The highest macro F_1 score (0.73) was obtained using RoBERTa with T5 concatenation. The BioBERT model outperformed BERT, ALBERT and all other domain-specific models for all representations except *concatenation* where BERT yielded a slightly higher overall F_1 score. Surprisingly, TAPAS achieved the lowest score. We believe that this is attributed to the small dataset; while TAPAS is one of the best-performing models on TabFact (Eisenschlos et al., 2020), our training set is much smaller, which can pose a challenge to the BERT-based model.

Performance on refuted claims. Across all applied models and table representations, we obtained a noticeable low F_1 score for PubHealthTab refuted claims compared to the two other veracity classes, support and NEI. The F_1 scores ranged from 0 (ClinicalBERT with concatenation) to 0.52 (RoBERTa and T5 concatenation).

To determine if this scenario was specific to our dataset, we compared the F_1 scores we obtained on our dataset using RoBERTa with other table fact-checking datasets. The results are shown in Table 8. While the F_1 score for PubHealthTab refuted claims was between 0.39 and 0.52 using RoBERTa, this value was between 0.65 and 0.87 for refuted claims from TabFact, InfoTabs and FEVEROUS. Whilst the low performance of RoBERTa on FEVEROUS NEI claims can be attributed to the imbalanced class distribution (Aly et al., 2021), this is not the case for PubHealthTab as the three veracity classes $\{$support, refute, NEI$\}$ are present in a ratio of 2:1:1 in our training set. We believe that the comparably low performance of RoBERTa on PubHealthTab *refute* claims is due to the fact that state-of-the-art representation and modelling approaches were previously evaluated on Wikipedia evidence tables. These approaches seem to struggle with noisy web tables: lacking table captions and headers, a higher ratio of numerical content, and a lower ratio of string-only content (see Section 4.4) could pose a challenge for generating table representations and for pre-trained models previously evaluated on tables from single data sources.

The results we obtained using RoBERTa on TabFact are lower compared to the other datasets. Whilst Chen et al. (2020b) do not report the results per class, the overall F_1 score we obtained is comparable to their baseline.

6 Conclusion

We introduced PubHealthTab, a table-based dataset for evidence-based fact checking centred on real-world public health claims. Our dataset comprises $1,942$ claim-table pairs, with tabular evidence data extracted from websites similar to those used by fact checkers. We described the dataset creation process and the steps taken to minimise biases and correlations. We evaluated state-of-the-art representation and modelling approaches and showed that the RoBERTa model achieves the highest performance on PubHealthTab across all representation methods compared to other models. In contrast to previous table-based fact-checking datasets that contain tables from single data sources, state-of-the-art models struggle to correctly classify refute claims from PubHealthTab against evidence consisting of mostly numerical data or with noisy text headers, making PubHealthTab a challenging dataset for table-based fact-checking research.

Ethics Statement

The PubHealthTab dataset can be used for developing and evaluating fact checking systems intended for a real-world context. The labels *supports*, *refutes* and *not enough information* describe a claim's veracity given the evidence table. We do not make any statement on PubHealthTab claims' truthfulness in a real-world context.

We obtained ethical clearance prior to crowdsourcing from the relevant authority in the academic institution. We informed the participants about the data being collected and its purpose. Participants had the opportunity to withdraw at any time and to provide feedback at the end of each task. All workers were from English speaking countries. The payment was above the minimum wage and decided based on the time workers spent on the pilot tasks. For the first and third tasks we paid $0.75USD$ (2.5 minutes per task on average) and for the second $1.35USD$ (average 5 minutes per task).

Acknowledgements

The authors acknowledge support from the Distributed AI (DAI) research group at King's College London for creating the dataset.

References

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. FEVEROUS: fact extraction and verification over unstructured and structured information. *CoRR*, abs/2106.05707.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. 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 4685–4697, Hong Kong, China. Association for Computational Linguistics.

Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2015. Soylent: a word processor with a crowd inside. *Communications of the ACM*, 58(8):85–94.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Yang Wang, and William W. Cohen. 2021. Open question answering over tables and text. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang. 2020a. Logical natural language generation from open-domain tables. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7929–7942, Online. Association for Computational Linguistics.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020b. TabFact: A large-scale dataset for table-based fact verification. In *8th International Conference on Learning Representations, ICLR*. OpenReview.net.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020c. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas Diggelmann, Jordan L. Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. CLIMATE-FEVER: A dataset for verification of real-world climate claims. *CoRR*, abs/2012.00614.

Shayan Doroudi, Ece Kamar, Emma Brunskill, and Eric Horvitz. 2016. Toward a learning science for complex crowdsourcing tasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2623–2634. ACM.

Julian Eisenschlos, Syrine Krichene, and Thomas Müller. 2020. Understanding tables with intermediate pre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 281–296, Online. Association for Computational Linguistics.

Stefan Evert. 2005. The statistics of word cooccurrences: word pairs and collocations.

Alvan R Feinstein and Domenic V Cicchetti. 1990. High agreement but low kappa: I. the problems of two paradoxes. *Journal of clinical epidemiology*, 43(6):543–549.

Ujwal Gadiraju, Besnik Fetahu, and Ricardo Kawase. 2015. Training workers for improving performance in crowdsourcing microtasks. In *Design for Teaching and Learning in a Networked World - 10th European Conference on Technology Enhanced Learning, EC-TEL*, volume 9307 of *Lecture Notes in Computer Science*, pages 100–114. Springer.

Vivek Gupta, Maitrey Mehta, Pegah Nokhiz, and Vivek Srikumar. 2020. INFOTABS: Inference on tables as semi-structured data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2309–2324, Online. Association for Computational Linguistics.

Darryl Hannan, Akshay Jain, and Mohit Bansal. 2020. ManyModalQA: Modality disambiguation and QA over diverse inputs. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI*, pages 7879–7886. AAAI Press.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. TaPas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333, Online. Association for Computational Linguistics.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. ClinicalBERT: Modeling clinical notes and predicting hospital readmission. *CoRR*, abs/1904.05342.

Mohit Iyyer, Wen-tau Yih, and Ming-Wei Chang. 2017. Search-based neural structured learning for sequential question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1821–1831, Vancouver, Canada. Association for Computational Linguistics.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR*. OpenReview.net.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Belinda Z. Li, Sewon Min, Srinivasan Iyer, Yashar Mehdad, and Wen-tau Yih. 2020. Efficient one-pass end-to-end entity linking for questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6433–6441, Online. Association for Computational Linguistics.

Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Nikolay Babulkov, Alex Nikolov, Gautam Kishore Shahi, Julia Maria Struß, and Thomas Mandl. 2021. The CLEF-2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR*, volume 12657 of *Lecture Notes in Computer Science*, pages 639–649. Springer.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Jeppe Nørregaard and Leon Derczynski. 2021. DanFEVER: claim verification dataset for Danish. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 422–428, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language*

Processing (Volume 1: Long Papers), pages 1470–1480, Beijing, China. Association for Computational Linguistics.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:140:1–140:67.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. 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 3419–3425, Hong Kong, China. Association for Computational Linguistics.

Lya Hulliyyatus Suadaa, Hidetaka Kamigaito, Kotaro Funakoshi, Manabu Okumura, and Hiroya Takamura. 2021. Towards table-to-text generation with numerical reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1451–1465, Online. Association for Computational Linguistics.

James Thorne, Max Glockner, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2021. Evidence-based verification for real world information needs. *CoRR*, abs/2104.00640.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. 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 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Nancy X. R. Wang, Diwakar Mahajan, Marina Danilevsky, and Sara Rosenthal. 2021. SemEval-2021 task 9: Fact verification and evidence finding for tabular data in scientific documents (SEM-TAB-FACTS). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 317–326, Online. Association for Computational Linguistics.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2SQL: Generating structured queries from natural language using reinforcement learning. *CoRR*, abs/1709.00103.

A Supplementary Materials

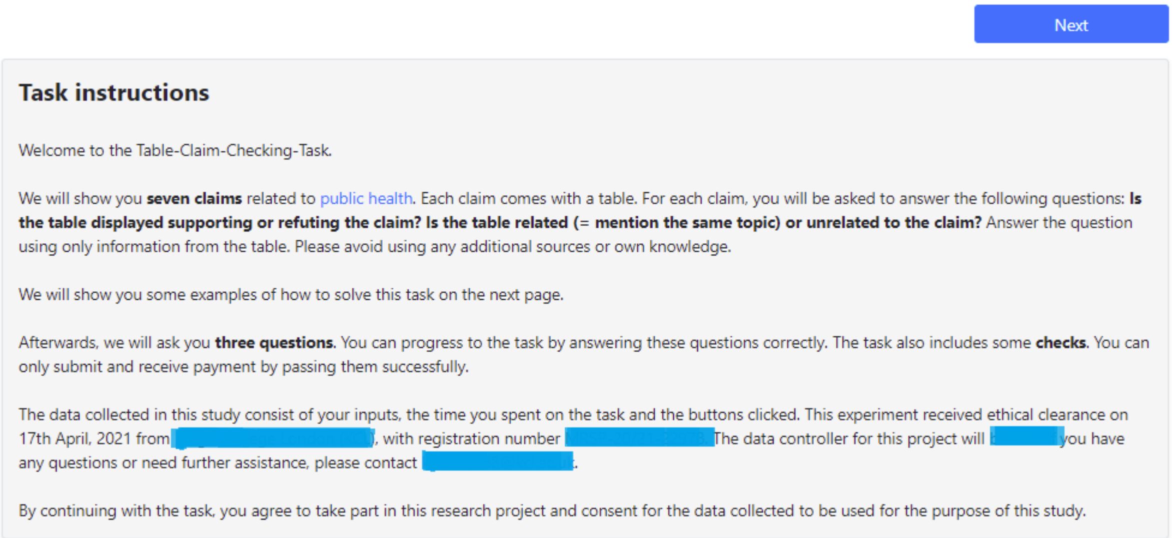
A.1 Dataset Creation

We evaluated the following conditions for the second crowdsourcing task. Workers could only submit their work if all checks were passed:

- A veracity label is selected for the adjusted claim.
- Minimum 2.5 seconds are spend on each HIT page for adjusting the claim.
- Adjusted claim length is between 5 and 30 tokens.
- The adjusted claim is different from the initial claim.
- The adjusted claim text does not contain ambiguous words, i.e. *maybe, probably, mostly, occasionally, frequently, might, many, few, some, several, most of, sometimes*.
- The adjusted claim does not contain negation words, i.e. *not, never, none, nobody*.

A.2 Experiments

After hyperparameter tuning on the validation set, we selected the following parameters for the different modelling approaches displayed in Table 9.



Next

Task instructions

Welcome to the Table-Claim-Checking-Task.

We will show you **seven claims** related to public health. Each claim comes with a table. For each claim, you will be asked to answer the following questions: **Is the table displayed supporting or refuting the claim? Is the table related (= mention the same topic) or unrelated to the claim?** Answer the question using only information from the table. Please avoid using any additional sources or own knowledge.

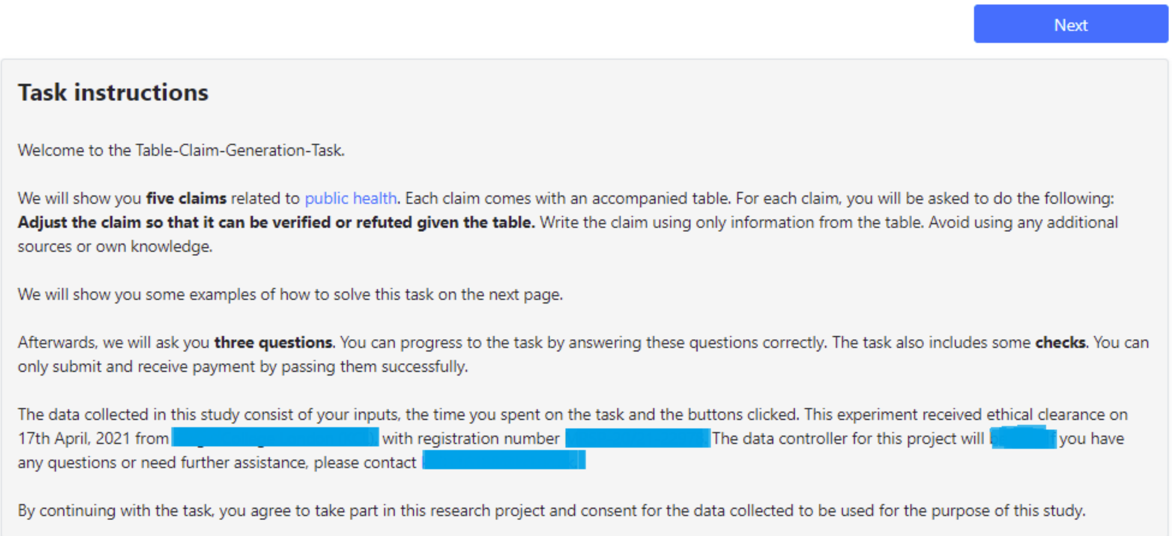
We will show you some examples of how to solve this task on the next page.

Afterwards, we will ask you **three questions**. You can progress to the task by answering these questions correctly. The task also includes some **checks**. You can only submit and receive payment by passing them successfully.

The data collected in this study consist of your inputs, the time you spent on the task and the buttons clicked. This experiment received ethical clearance on 17th April, 2021 from [redacted], with registration number [redacted] The data controller for this project will [redacted] you have any questions or need further assistance, please contact [redacted].

By continuing with the task, you agree to take part in this research project and consent for the data collected to be used for the purpose of this study.

Figure 3: Introduction text for *table relevance* and *verification* task.



Next

Task instructions

Welcome to the Table-Claim-Generation-Task.

We will show you **five claims** related to public health. Each claim comes with an accompanied table. For each claim, you will be asked to do the following: **Adjust the claim so that it can be verified or refuted given the table.** Write the claim using only information from the table. Avoid using any additional sources or own knowledge.

We will show you some examples of how to solve this task on the next page.

Afterwards, we will ask you **three questions**. You can progress to the task by answering these questions correctly. The task also includes some **checks**. You can only submit and receive payment by passing them successfully.

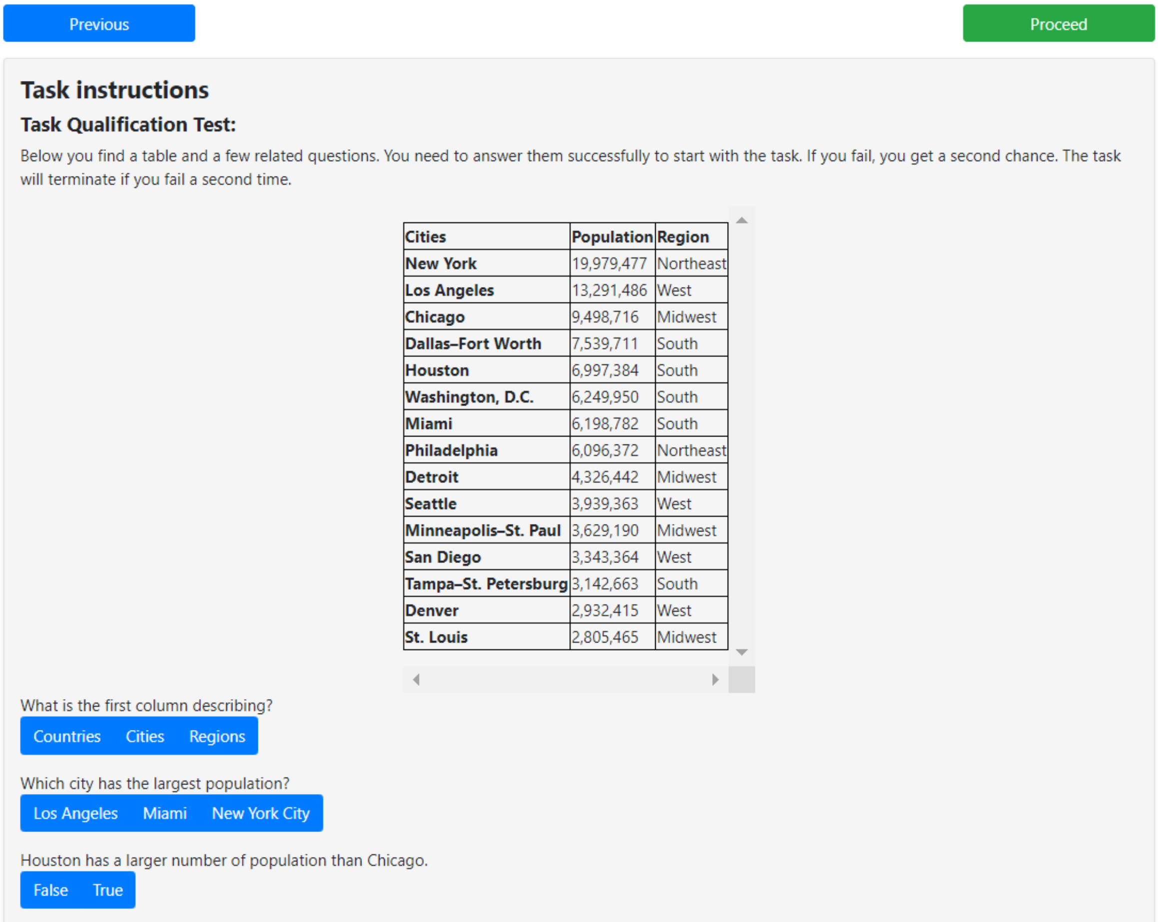
The data collected in this study consist of your inputs, the time you spent on the task and the buttons clicked. This experiment received ethical clearance on 17th April, 2021 from [redacted] with registration number [redacted] The data controller for this project will [redacted] you have any questions or need further assistance, please contact [redacted]

By continuing with the task, you agree to take part in this research project and consent for the data collected to be used for the purpose of this study.

Figure 4: Introduction text for *claim adjustment* task.

Model	TE	BS	LR	WD
BERT	5	4	1e-5	0.001
AlBERT	5	16	1e-5	0.001
RoBERTa	4	8	1e-5	0.01
BioBERT	5	4	1e-5	0.001
BlueBERT	5	8	1e-5	0.001
ClinicalBERT	4	4	1e-5	0.01

Table 9: Hyperparameters evaluated on the PubHealthTab dataset: training epochs (TE), batch size (BS), learning rate (LR), weight decay (WD).



Previous

Proceed

Task instructions

Task Qualification Test:

Below you find a table and a few related questions. You need to answer them successfully to start with the task. If you fail, you get a second chance. The task will terminate if you fail a second time.

Cities	Population	Region
New York	19,979,477	Northeast
Los Angeles	13,291,486	West
Chicago	9,498,716	Midwest
Dallas–Fort Worth	7,539,711	South
Houston	6,997,384	South
Washington, D.C.	6,249,950	South
Miami	6,198,782	South
Philadelphia	6,096,372	Northeast
Detroit	4,326,442	Midwest
Seattle	3,939,363	West
Minneapolis–St. Paul	3,629,190	Midwest
San Diego	3,343,364	West
Tampa–St. Petersburg	3,142,663	South
Denver	2,932,415	West
St. Louis	2,805,465	Midwest

What is the first column describing?

Countries Cities Regions

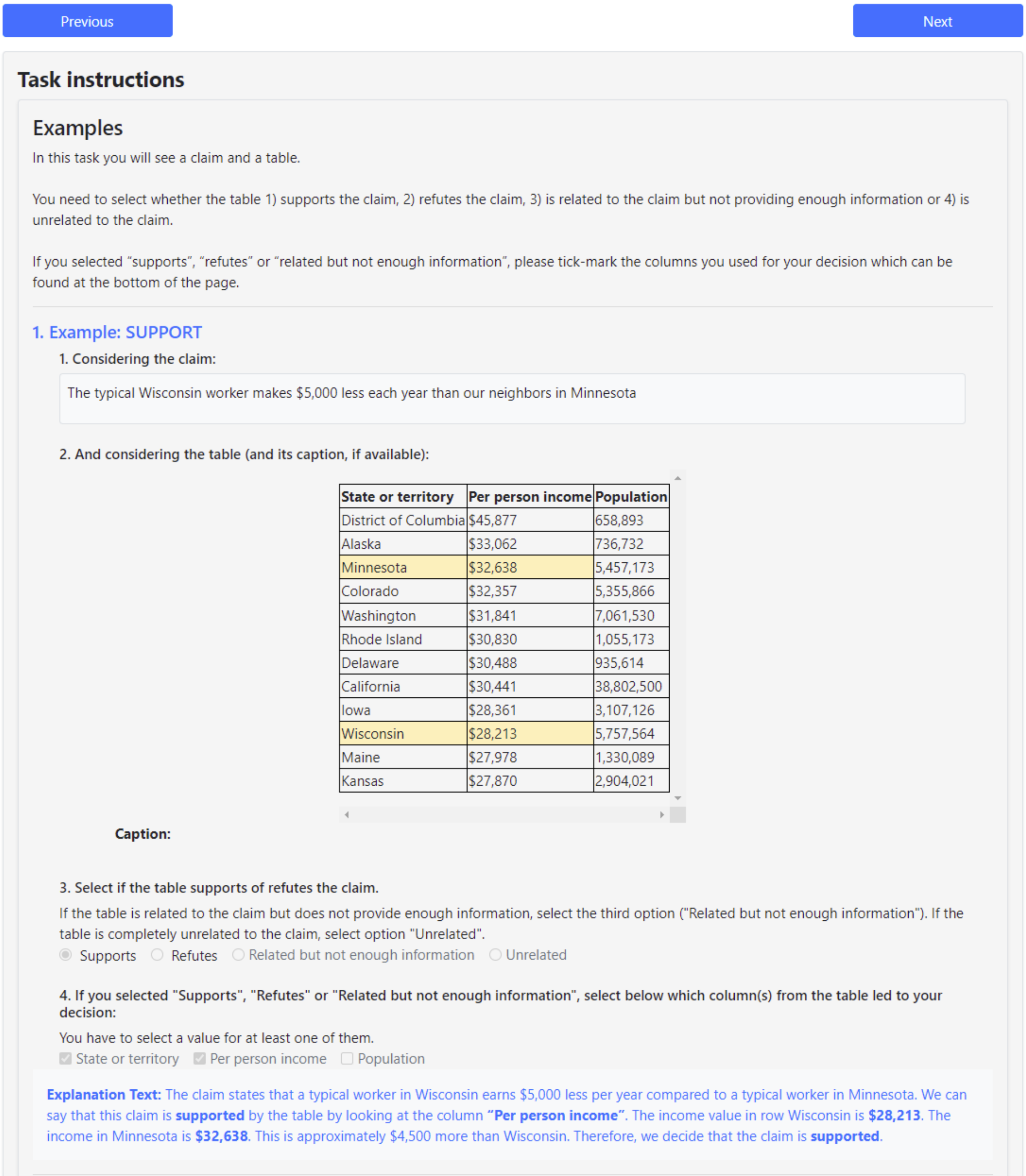
Which city has the largest population?

Los Angeles Miami New York City

Houston has a larger number of population than Chicago.

False True

Figure 5: Crowdsourcing qualification test.



Previous
Next

Task instructions

Examples

In this task you will see a claim and a table.

You need to select whether the table 1) supports the claim, 2) refutes the claim, 3) is related to the claim but not providing enough information or 4) is unrelated to the claim.

If you selected “supports”, “refutes” or “related but not enough information”, please tick-mark the columns you used for your decision which can be found at the bottom of the page.

1. Example: SUPPORT

1. Considering the claim:

The typical Wisconsin worker makes $5,000 less each year than our neighbors in Minnesota

2. And considering the table (and its caption, if available):

State or territory	Per person income	Population
District of Columbia	$45,877	658,893
Alaska	$33,062	736,732
Minnesota	$32,638	5,457,173
Colorado	$32,357	5,355,866
Washington	$31,841	7,061,530
Rhode Island	$30,830	1,055,173
Delaware	$30,488	935,614
California	$30,441	38,802,500
Iowa	$28,361	3,107,126
Wisconsin	$28,213	5,757,564
Maine	$27,978	1,330,089
Kansas	$27,870	2,904,021

Caption:

3. Select if the table supports of refutes the claim.

If the table is related to the claim but does not provide enough information, select the third option ("Related but not enough information"). If the table is completely unrelated to the claim, select option "Unrelated".

Supports Refutes Related but not enough information Unrelated

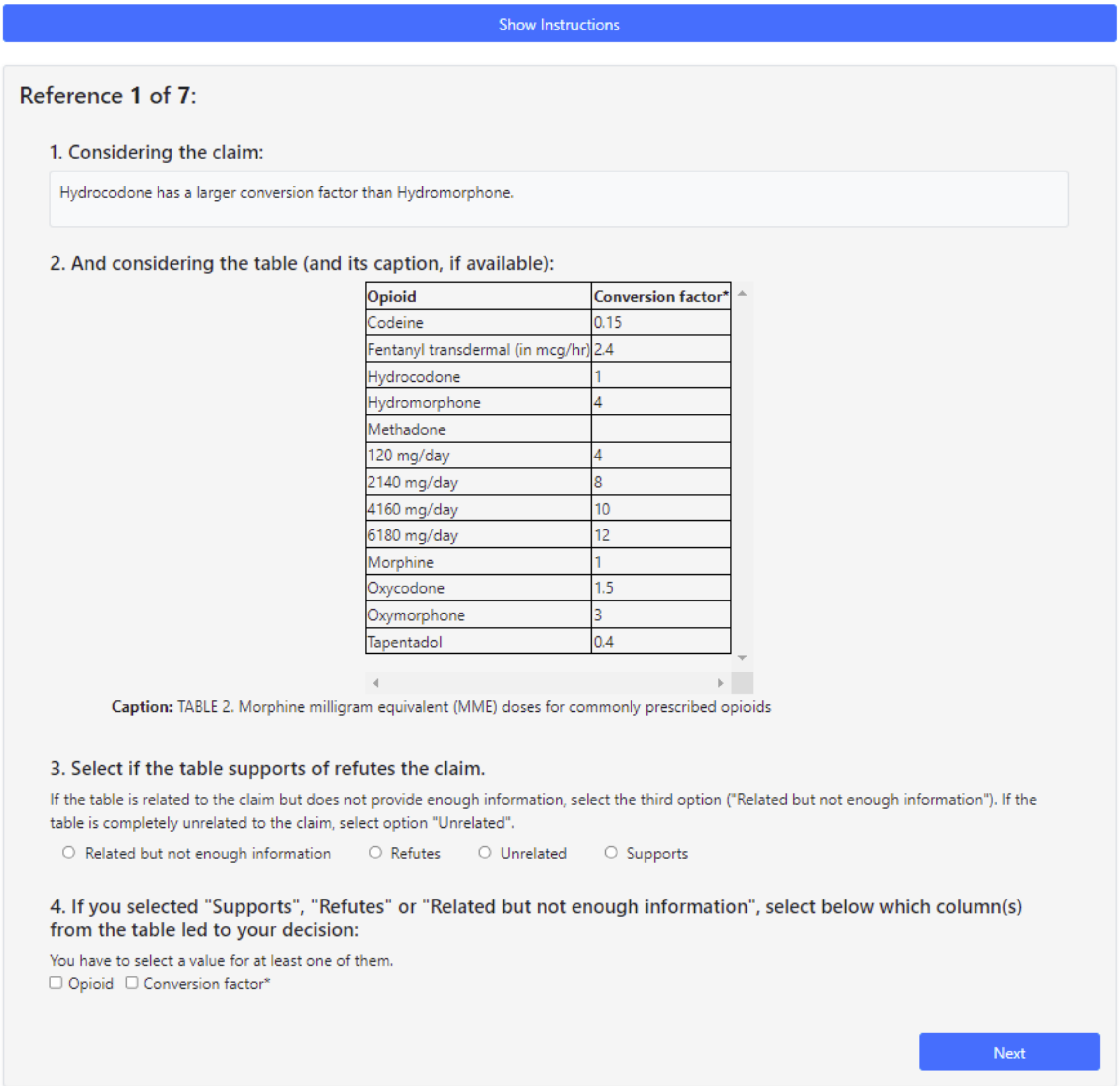
4. If you selected "Supports", "Refutes" or "Related but not enough information", select below which column(s) from the table led to your decision:

You have to select a value for at least one of them.

State or territory Per person income Population

Explanation Text: The claim states that a typical worker in Wisconsin earns $5,000 less per year compared to a typical worker in Minnesota. We can say that this claim is **supported** by the table by looking at the column **“Per person income”**. The income value in row Wisconsin is **$28,213**. The income in Minnesota is **$32,638**. This is approximately $4,500 more than Wisconsin. Therefore, we decide that the claim is **supported**.

Figure 6: Author-annotated crowdsourcing example.



Show Instructions

Reference 1 of 7:

1. Considering the claim:

Hydrocodone has a larger conversion factor than Hydromorphone.

2. And considering the table (and its caption, if available):

Opioid	Conversion factor*
Codeine	0.15
Fentanyl transdermal (in mcg/hr)	2.4
Hydrocodone	1
Hydromorphone	4
Methadone	
120 mg/day	4
2140 mg/day	8
4160 mg/day	10
6180 mg/day	12
Morphine	1
Oxycodone	1.5
Oxymorphone	3
Tapentadol	0.4

Caption: TABLE 2. Morphine milligram equivalent (MME) doses for commonly prescribed opioids

3. Select if the table supports of refutes the claim.

If the table is related to the claim but does not provide enough information, select the third option ("Related but not enough information"). If the table is completely unrelated to the claim, select option "Unrelated".

○ Related but not enough information ○ Refutes ○ Unrelated ○ Supports

4. If you selected "Supports", "Refutes" or "Related but not enough information", select below which column(s) from the table led to your decision:

You have to select a value for at least one of them.

☐ Opioid ☐ Conversion factor*

Next

Figure 7: User Interface for the *table relevance* and *verification* task.

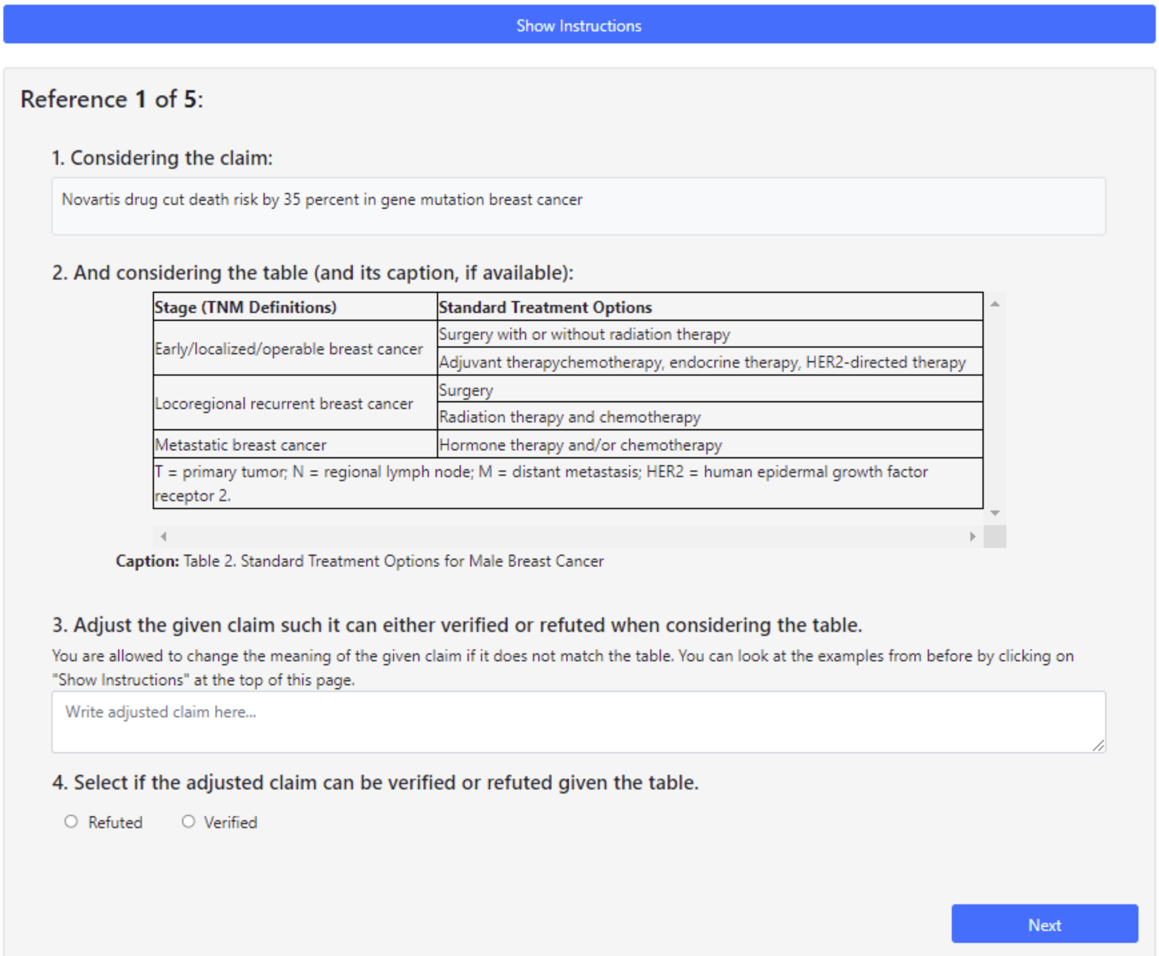


Figure 8: User Interface for the *claim adjustment* task.