Intra-Template Entity Compatibility based Slot-Filling for Clinical Trial Information Extraction

Christian Witte
Bielefeld University, Germany
cwitte@
techfak.uni-bielefeld.de

Philipp Cimiano
Bielefeld University, Germany
cimiano@
cit-ec.uni-bielefeld.de

Abstract

We present a deep learning based information extraction system that can extract the design and results of a published abstract describing a Randomized Controlled Trial (RCT). In contrast to other approaches, our system does not regard the PICO elements as flat objects or labels but as structured objects. We thus model the task as the one of filling a set of templates and slots; our two-step approach recognizes relevant slot candidates as a first step and assigns them to a corresponding template as second step, relying on a learned pairwise scoring function that models the compatibility of the different slot values. We evaluate the approach on a dataset of 211 manually annotated abstracts for Type 2 Diabetes and Glaucoma, showing the positive impact of modelling intra-template entity compatibility. As main benefit, our approach yields a structured object for every RCT abstract that supports the aggregation and summarization of clinical trial results across published studies and can facilitate the task of creating a systematic review or meta-analysis.

1 Introduction

The evidence based medicine (EBM) paradigm (Sackett et al., 1996) propagates that individual medical decisions are taken on the basis of the best available clinical evidence. The activity of summarizing the existing body of evidence is a core activity to support EBM and its most prominent instrument is the systematic review. Creating a systematic review involves a high effort, involving on average 67.3 weeks and involving 5 authors per review on average (Borah et al., 2017). Keeping systematic reviews up to date involves an even much higher and continuous effort (Koch, 2006; Beller et al., 2013).

Thus, there is increased interest in partially automatizing the creation of systematic reviews (O'Connor et al., 2019). A significant hindrance for the automation of systematic reviews is that data needs to be extracted by hand from published studies. This problem could be alleviated if publications were machine readable, or could be turned into a structured, machine readable form by information extraction methods (Liu et al., 2016; Wu et al., 2020).

The methods that so far have been applied to the automatic extraction of information from clinical trial publications follow the PICO framework and attempt to extract the Population, Intervention, Comparator and Outcomes from a publication. Most approaches formalize the task as a tagging or classification problem. Some approaches for instance attempt to tag spans in the text and label them with the PICO elements (e.g. (Trenta et al., 2015)). Others classify complete text segments into these classes (Boudin et al., 2010; Jin and Szolovits, 2018).

However, the PICO elements denote structured objects rather than plain tags or classes. An intervention is described by a drug, frequency of administration, administration route, dose, etc. An outcome is described by a certain increase or decrease of a value from a baseline condition, refers to a certain primary or secondary endpoint, and there are outcomes for each arm of a trial that need to be compared to each other. In spite of being structured objects, most previous work treats these elements as flat and unstructured. Treating them as such makes the automatic aggregation and summarization of results challenging if not impossible.

Towards treating information extraction from clinical publications as a problem of predicting structured elements, we model the task as a template extraction task in which each template consists of a number of slots to be extracted. In Table 1 we provide an overview of all the templates we consider in this work and the number and types of slots they have.

Towards extracting these templates and thus a

Proceedings of the BioNLP 2022 workshop, Dublin, Ireland, pages 178–192

structured representation of a clinical trial and its results, we present a novel deep learning architecture. The architecture first labels spans of text as candidate slot fillers of a particular slot in a first step. In a second step, the filler is assigned to an instance of a template. With this two-step architecture, we can transform each clinical trial abstract into a structured representation that supports down-stream aggregation of results.

As there can be multiple interventions, arms and outcomes in a given study, an important challenge is to predict how many instances of each template occur in a given clinical trial publication. We leave this subpart of the problem for future work and assume that the number of interventions, arms and outcomes is known a priori. This assumption is reasonable as this information is typically contained in existing registries for trials such as https://www.clinicaltrials.gov/.

When assigning slot fillers to templates, it is important to model the dependencies between the different slots as some values might be compatible while others not. We model this compatibility by a trained function that predicts a compatibility score.

In summary our contributions are as follows:

- We propose a new approach to extracting evidence from clinical trial publications that consists in instantiating a set of pre-defined templates. As a result, the key findings of a clinical trial can be represented in a fully structured and machine-readable form that supports down-stream aggregation. To the best of out knowledge, we present the first template-filling IE approach in the clinical trial domain.

- We present a novel two-step deep learning based architecture that first recognizes slot candidates and then assigns these candidates to instances of templates. At a second step, candidates for slot fillers are assigned to a template instance.

- We show that it is possible to extract fine grained candidates of slot fillers from 37 classes yielding very good results of micro $F_1 = 76.21\%$ on the Glaucoma and $F_1 = 76.49\%$ the Type 2 Diabetes Mellitus (T2DM) dataset (Sanchez-Graillet et al., 2021), respectively.

- We introduce an intra-template entity compatibility optimization procedure for distributing entities to template instance of the same type. We show the impact of including a function for scoring the compatibility of slot assignments, and show that it improves extraction results in terms of F-Measure by 6.34% and 3.95% on the Glaucoma and T2DM dataset, respectively.

2 Related Work

The template extraction and slot filling task we address is related to the field of event extraction (Frisoni et al., 2021) where the goal is to extract so called *event triggers* and the arguments of the events. Our templates can be seen as complex events and our slots as arguments thereof.

Wang et al. (2020) adopt the question answering paradigm to extract events from biomedical texts. They introduce two different types of questions for extracting event triggers and event arguments. However, in their approach the extraction of event arguments also relies on the extraction of event triggers.

Adel et al. (2018) introduce a framework for task-independent template-based information extraction. Their approach first identifies text spans representing slot-fillers as in our approach. However, their system relies on the successful identification of anchor spans representing template instances as they cast the assignment of slot-fillers to template instances as a binary classification between anchor spans and other text spans. The slot filling system proposed by Zhang et al. (2017) is a neural architecture that can exploit the combination of semantic similarity-based attention and position-based attention. The authors address a relation extraction task and develop a large corpus of annotated relations, TACRED (Zhang et al., 2017).

More recent work has framed the task of relation extraction in the biomedical field as a slot filling task as well (Papanikolaou and Bennett, 2021). However, the work is limited to extracting binary relationships (drug-drug, compound-drug and compound-disease).

Early work on extracting information from text describing clinical trials has focused on the classification of sentences into sections of papers describing Randomized Controlled Trials (RCTs), e.g. Methods, Results, etc. (McKnight and Srinivasan, 2003; Hirohata et al., 2008; Chung, 2009). Such systems tackle a very coarse-grained information

Template Name	#Slots	Slots
Arm	7	AdverseEffect, FinalNumPatientsArm, Intervention, NumPatientsLeftArm, NumberPatientsArm, Outcome, RelFinalNumPatientsArm,
ClinicalTrial	15	analysesHealthCondition, AllocationRatio, AnalysisApproach, Arm, CTDesign, CTduration, ConclusionComment, DiffBetweenGroups, EvidQualityIndicator, FinalNumberPatientsCT, NumPatientsLeftCT, NumberPatientsCT, ObjectiveDescription, Population, RelNumPatientsLeftCT
DiffBetweenGroups	8	ConfIntervalDiff, DiffGroupAbsValue, DiffGroupRelValue, Outcome1, Outcome2, PvalueDiff, StandardDevDiff, StandardErrorDiff
Endpoint	4	AggregationMethod, BaselineUnit, EndPointDescription, MeasurementDevice
Intervention	5	Duration, Frequency, Interval, Medication, RelativeFreqTime
Medication	6	ApplicationCondition, DeliveryMethod, DoseDescription, DoseUnit, DoseValue, Drug
Outcome	26	BaselineValue, ChangeValue, ConfIntervalBL, ConfIntervalChangeValue, ConfIntervalNumAffected, ConfIntervalResValue, Endpoint, NumberAffected, ObservedResult, PValueBL, PValueChangeValue, PValueNumAffected, PValueResValue, PercentageAffected, RelativeChangeValue, ResultMeasuredValue, SdDevBL, SdDevChangeValue, SdDevNumAffected, SdDevResValue, SdErrorBL, SdErrorChangeValue, SdErrorNumAffected, SdErrorResValue, SubGroupDescription, TimePoint
Population	7	in AvgAge, Country, Ethnicity, Gender, MaxAge, MinAge, Precondition
Publication	6	describes, Author, Journal, PMID, PublicationYear, Title

Table 1: Template types and corresponding slots

extraction task as they do not extract the actual content or results of a published RCT, but only extract correspondences between content and the standard sections used to describe a clinical trial in a publication. Such a sentence classification task can support the indexation and thus retrieval of information from a published RCT, but does not support the use case we consider, i.e. the aggregation of evidence across published trials.

Beyond the classification of sentences into sections of an article, other authors have considered the classification of sentences into PICO elements, that is classifying a sentence in a published clinical trial with respect to whether it describes the Population, Intervention, Comparator or an Outcome (Demner-Fushman and Lin, 2007; Chung, 2009; Boudin et al., 2010; Jin and Szolovits, 2018). Such approaches are able to extract information at a more detailed granularity, but they still do not support aggregation of evidence across studies as the mere classification of sentences with respect to PICO elements does not provide a semantic structure that can be used to describe the key results of a study.

The work by Trenta et al. (2015) goes one step further in that it tags spans of text in an RCT abstract into the PICO classes, considering the following classes: patient group, intervention, arm, control arm, measured outcome, etc. Trenta et al. (2015) rely on maximum entropy models and use integer linear programming to define constraints on the classified tokens, e.g., such that *Results* can not occur in the *Methods* section. They show that their approach is able to extract evidence tables from RCT abstracts. Yet, the different spans extracted are only indirectly related to each other in the model of Trenta et al. (2015). This gap is addressed by the approach of Nye et al. (2020), which beyond extracting PICO elements (intervention arms, outcome measures, results) also relates the different snippets to each other, yielding a relational structure.

Inspired by the work of Trenta et al. (2015) as well as Nye et al. (2020) we go one step further in extracting a complete structured object from an RCT abstract comprising of nine main template types with overall 85 slots. To our knowledge, this is thus the most fine-grained representation that so far has been considered by an information extraction system in the clinical domain.

3 Model

As already mentioned in the introduction, our proposed model consists of a two-step architecture. The first component, the entity extraction (EE) module, identifies spans of slot filler candidates (SFCs). We assume that we have a set of template types $\mathcal{T} = t_1, ..., t_{|\mathcal{L}|}$ which correspond to the template types depicted in Table 1, where $\mathcal{L}$ denotes the number of template types. We refer to the slot j of template t_i as $s_{i,j}$. The set of all slots is $\mathcal{S} = \bigcup_{i,j}\{s_{i,j}\}$ and the set of slots of template type t is $\mathcal{S}_t = \bigcup_j\{s_{t,j}\}$.

The set of all SFCs extracted within an abstract is denoted by $\mathcal{E}$. Formally speaking, the entity

extraction module implements a function f_{EE} that maps each slot filler candidate into a slot type, i.e. $f_{EE} : \mathcal{E} \rightarrow \mathcal{S}$.

The second component, the template assignment (TA) module, maps each slot filler to a particular instance of a template. Hereby, we can have multiple instances of a given template type. For instance, in the general case a clinical study might describe multiple interventions, multiple endpoints and multiple outcomes. We denote the i-th instance of template t by $T_i^{(t)}$. The set of all template instances is thus $\theta = \bigcup\{T_i^{(t)}\}$ and the number of template instances of template type t is denoted by m_t. The second component thus realizes a function $f_{TA} : \mathcal{E} \rightarrow \theta$. We denote the template type to which SFC e_j has been assigned to as y_{e_j}.

Take the following sentence as an example: *Mean 24-h IOP with BTFC was significantly lower than with latanoprost (18.9 vs 21.2 mmHg; p < 0.001)*. The first component would recognize the spans *18.9* and *21.2* and map them both to the slot type `ResultMeasuredValue`. Then the TA module assignes these identifies SFCs to template instances of type Outcome, together with other SCFs extracted from other sentences.

Note that both modules fully specify a mapping from entities detected in the clinical trial abstract to fully instantiated templates, where f_{EE} identifies and classifies text spans into slots and f_{EA} identifies the appropriate instance of a template.

We describe both modules in more detail subsequently. In particular, as the assignment of text spans to slots and template instances should not be modelled completely independently, we introduce an additional component that computes an overall score for a given template instance that quantifies the compatibility of the assigned text spans to all of the slots of the template instance. These scores can be regarded as factors as used in factor graph models (Kschischang et al., 2001). In order to reduce the complexity, we model the interaction between different slots in a pairwise fashion, limiting the scope of these factors to two slots.

3.1 Entity Extraction Module

The entity extraction module identifies token spans in the input document which either represent named entities or literals. The extracted token spans are later assigned to slots by the module described in section 3.2. We represent documents $\mathcal{D}$ by a sequence of sentences $(s_1, \ldots, s_{n_S})$ where each sentence s_i in turn is represented by a sequence of tokens $(w_1^{(s_i)}, \ldots, w_{n_{s_i}}^{(s_i)})$, where n_S denotes the number of sentences in document $\mathcal{D}$ and n_{s_i} denotes the number of tokens of sentence s_i. We adopt the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) architecture for computing contextualized token representations within the input document. A BERT layer is a stack of K identical Transformers (Vaswani et al., 2017) which captures pairwise token dependencies via an attention mechanism. Since most BERT implementations limit the length of input sequences by k_{max}, we split the sequence of sentences of the input document into n_C subsequences (chunks) if the number of tokens of the document exceeds this upper bound. We use the special token $[SEP]$ to separate sentences within a given chunk c_i and prepend the special token $[CLS]$ to each chunk which allows for capturing global context information for each chunk. The output for chunk c_i of the K-th Transformer of the BERT layer is a sequence of contextualized vectors $\mathbf{h}_1^{(c_i)}, \ldots, \mathbf{h}_{n_{c_i}}^{(c_i)} \in R^{d_{bert}}$, where the vector $\mathbf{h}_j^{(c_i)}$ represents the j-th token of chunk c_i, d_{bert} denotes the dimension of the BERT model and n_{c_i} denotes the number of tokens in chunk c_i.

Entity extraction is implemented through two dense layers which independently predict which tokens are start and/or end positions of entities which are referenced by a slot. This is achieved by using the set of slots $\mathcal{S}$ as entity types. Then the predicted entity type indirectly specifies the type of the template the entity has to be assigned to since no pair of template types shares the same set of slots. More formally, the two dense layers are given by

$$\hat{\mathbf{y}}_{j,start}^{(c_i)} = softmax(\mathbf{W}_{start}\mathbf{h}_j^{(c_i)} + \mathbf{b}_{start}) \quad (1)$$

$$\hat{\mathbf{y}}_{j,end}^{(c_i)} = softmax(\mathbf{W}_{end}\mathbf{h}_j^{(c_i)} + \mathbf{b}_{end}) \quad (2)$$

where $\mathbf{W}_{start}, \mathbf{W}_{end} \in R^{(|\mathcal{S}|+1)\times d_{bert}}$, $\mathbf{b}_{start}, \mathbf{b}_{end}, \in R^{d_{bert}}$.

The prediction of the slot is performed as follows:

$$\begin{aligned} \hat{y}_{j,start}^{(c_i)} &= \arg\max \hat{\mathbf{y}}_{j,start}^{(c_i)} \\ \hat{y}_{j,end}^{(c_i)} &= \arg\max \hat{\mathbf{y}}_{j,end}^{(c_i)} \end{aligned}$$

At inference time we join the predicted start and end positions by assigning the closest predicted end

position p_{end} of type t within the same sentence to each predicted start position p_{start} of type t under the constraint $p_{start} <= p_{end}$.

Finally we compute a vector representation $\mathbf{e}_k$ for each extracted SFC e_k by summing the vectors $\mathbf{h}_j^{c_i}$ of the corresponding start and end tokens of the SFC, followed by a dense layer with a ReLu activation function (Agarap, 2018).

3.2 Template Assignment Module

The TA module described in this section assigns each SFC $e_j \in \mathcal{E}$ extracted by the entity extraction module to a template in θ. As we know the slot y_{e_j} that e_j has been assigned to, the template type t of y_{e_j} determines the subset θ_t of template instances in the set θ that e_j can be assigned to. This reduces the search space considerably and essentially allows us to model the template assignment task as the one of inducing a partition.

Let's assume that SFCs are grouped into $|\mathcal{L}|$ disjoint subsets $\mathcal{E}_1, \ldots, \mathcal{E}_{|\mathcal{L}|}$ according to their type t, that is:

$$\mathcal{E}_t = \{e_j \in \mathcal{E} \mid y_{e_j} \in \mathcal{S}_t\}, \quad t \in \mathcal{L} \tag{3}$$

The task of template assignment can be reduced to the task of partitioning each set $\mathcal{E}_t$ into a partition $\mathcal{P}_t = \{\mathcal{T}_1^{(t)}, \ldots, \mathcal{T}_{m_t}^{(t)}\}$ of $\mathcal{E}_t$ where each set $\mathcal{T}_i^{(t)}$ contains the SFCs assigned to template instance $T_i^{(t)}$.

We call a partition $\mathcal{P}_t$ of the set $\mathcal{E}_t$ valid if each SFC $e_j \in \mathcal{E}_t$ is assigned to exactly one partition $\mathcal{T}_i^{(t)} \in \mathcal{P}_t$ and we denote the set of all valid partitions for the set $\mathcal{E}_t$ as $\mathcal{U}_t$.

We propose a pairwise intra-template entity compatibility optimization objective which measures the joint compatibility of the SFCs within the sets $\mathcal{T}_i^{(t)}$ of a partition. Let $q : \mathcal{E} \times \mathcal{E} \rightarrow [0, 1]$ denote the function which measures the compatibility between two SFCs e_j, e_k, where $q(e_j, e_k) = 1$ means maximal compatibility and $q(e_j, e_k) = 0$ means minimal compatibility. Note that we assume that q is symmetric in its arguments, i.e., $q(e_j, e_k) = q(e_k, e_j)$. Then the mean pairwise entity compatibility score $h(\mathcal{T}_i^{(t)})$ for the set $T_i^{(t)}$ is given by

$$h(\mathcal{T}_i^{(t)}) = \frac{1}{\frac{m_t!}{2(m_t-2)!}} \sum_{e_j, e_k \in \mathcal{T}_i^{(t)}, j<k} q(e_j, e_k) \tag{4}$$

and the compatibility score for partition $\mathcal{P}_t$ is the sum of the mean pairwise compatibility scores of each template set $\mathcal{T}_i^{(t)} \in \mathcal{P}_t$:

$$\sum_{\mathcal{T}_i^{(t)} \in \mathcal{P}_t} h(\mathcal{T}_i^{(t)}) \tag{5}$$

Given these definitions, we seek the partition $\hat{\mathcal{P}}_t \in \mathcal{U}_t$ which maximizes the compatibility score defined by Eq. (5). Hence the optimization problem proposed by our approach is given by

$$\hat{\mathcal{P}}_t = \arg\max_{\mathcal{P}_t \in \mathcal{U}_t} \sum_{\mathcal{T}_i^{(t)} \in \mathcal{P}_t} h(\mathcal{T}_i^{(t)}) \tag{6}$$

for all template types $t \in \mathcal{L}$. For arbitrary large entity sets $\mathcal{E}_t$, the sets $\mathcal{U}_t$ of valid partitions can become very large because of the combinatorial explosion, and hence finding the exact solution of the optimization problem defined by Eq. (6) can become intractable. Therefore we propose an approximate optimization method based on beam search which maintains a set $\mathcal{B}_t^{(z)}$ of n_B candidate solutions in each iteration z which are gradually refined. We define a candidate solution i for template type t as a pair $(\mathcal{E}_t^{(i)}, \mathcal{P}_t^{(i)})$, where $\mathcal{P}_t^{(i)}$ denotes the candidate partition and $\mathcal{E}_t^{(i)} \subseteq \mathcal{E}_t$ denotes the set of entities of that candidate solution which are not yet assigned to any template set $\mathcal{T}_i^{(t)} \in \mathcal{P}_t^{(i)}$. In each iteration z, we compute all successors of all candidate solutions $(\mathcal{E}_t^{(i)}, \mathcal{P}_t^{(i)}) \in \mathcal{B}_t^{(z)}$ by assigning an entity $e_j \in \mathcal{E}_t^{(i)}$ to a template set $\mathcal{T}_i^{(t)} \in \mathcal{P}_t^{(i)}$, which yields a set of new candidate solutions $\tilde{\mathcal{B}}_t^{(z)}$. Next we rank all candidate solutions in $\tilde{\mathcal{B}}_t^{(z)}$ by computing the mean intra-template entity compatibility score defined by Eq (5) for each candidate partition of the respective candidate solutions and keep only the best n_B ones, which yields the new beam $\mathcal{B}_t^{(z+1)}$ for the next iteration. After all entities for template type t have been assigned to a template after Z iterations, the partition $\mathcal{P}_t^{(i)}$ of the best ranked final candidate solution $(\mathcal{E}_t^{(i)}, \mathcal{P}_t^{(i)}) \in \mathcal{B}_t^{(Z)}$ is returned. The initial seed sets $\mathcal{B}_t^{(0)}$ of candidate solutions for each template type t are given by

$$\mathcal{B}_t^{(0)} = \{(\mathcal{E}_t, \{\mathcal{T}_i^{(t)}\}_{i=1}^{m_t})\}, \quad \mathcal{T}_i^{(t)} = \{\} \tag{7}$$

More details of the optimization procedure can be found in algorithm 1.

We implement the pairwise entity compatibility function $q(e_i, e_j)$ through summing the vector representations $\mathbf{e}_i$ and $\mathbf{e}_j$ of the corresponding entities

Data: Set of SFCs $\mathcal{E}$; entity compatibility function g; beam size n_B
Result: Partitions $\mathcal{P}_i, \ldots, \mathcal{P}_{|\mathcal{L}|}$
for $t \in \mathcal{L}$ **do**
  Compute set of SFCs $\mathcal{E}_t$ for template type t by Eq. (3)
  Compute beam seed set $\mathcal{B}_t^{(0)}$ defined by Eq. (7)
  $z \leftarrow 0$
  for $i \in \{1, \ldots, |\mathcal{E}_t|\}$ **do**
    Initialize set of successor candidate solutions $\tilde{\mathcal{B}}_t^{(z)}$ as empty set
    for $(\mathcal{E}_t^{(i)}, \mathcal{P}_t^{(i)}) \in \mathcal{B}_t^{(z)}$ **do**
      for $e_k \in \mathcal{E}_t^{(i)}$ **do**
        for $\mathcal{T}_j^{(t)} \in \mathcal{P}_t^{(i)}$ **do**
          Remove e_k from $\mathcal{E}_t^{(i)}$ which yields the set $\tilde{\mathcal{E}}_t^{(i)}$
          Add e_k to set $\mathcal{T}_j^{(t)}$ which yields $\tilde{\mathcal{T}}_j^{(t)}$
          Replace $\mathcal{T}_j^{(t)}$ in $\mathcal{P}_t^{(i)}$ by $\tilde{\mathcal{T}}_j^{(t)}$ which yields $\tilde{\mathcal{P}}_t^{(i)}$
          Add new candidate solution $(\tilde{\mathcal{E}}_t^{(i)}, \tilde{\mathcal{P}}_t^{(i)})$ to set $\tilde{\mathcal{B}}_t^{(z)}$
        end
      end
    end
    Rank all candidate solutions in $\tilde{\mathcal{B}}_t^{(z)}$ by Eq. (5)
    Keep the best ranked n_B candidate solutions from $\tilde{\mathcal{B}}_t^{(z)}$ which yields new batch $\tilde{\mathcal{B}}_t^{(z)}$
    $z \leftarrow z + 1$
  end
  Get best ranked candidate solution $(\hat{\mathcal{E}}_t^{(i)}, \hat{\mathcal{P}}_t^{(i)})$ from $\mathcal{B}_t^{(|\mathcal{E}_t|)}$
  $\mathcal{P}_t \leftarrow \hat{\mathcal{P}}_t^{(i)}$
end

Algorithm 1: Pseudo-code of our proposed approximate optimization method for maximizing the mean intra-template entity compatibility when assigning the extracted entities to templates

followed by a dense layer with sigmoid activation function. More formally:

$$\hat{q}(e_i, e_j) = \sigma(\mathbf{w}_{comp} \odot (\mathbf{e}_i + \mathbf{e}_j) + b_{comp}) \quad (8)$$

where $\mathbf{w}_{comp} \in R^{d_{bert}}$, $b_{comp} \in R$ and $\odot$ denotes the scalar product of two vectors.

3.3 Model Training

We train the model in end-to-end fashion by jointly minimizing the loss of the EE module and the TA module. The loss L_{EE} of the EE module is given by the cross entropy between the predicted SFC start position $\hat{\mathbf{y}}_{j,start}^{(c_i)}$ and ground truth SFC start position $\mathbf{y}_{j,start}^{(c_i)}$ plus the cross entropy between predicted SFC end positions $\hat{\mathbf{y}}_{j,end}^{(c_i)}$ and the ground truth SFC end positions $\mathbf{y}_{j,end}^{(c_i)}$.

The loss L_{TA} of the TA module is given by the cross entropy between the ground truth compatibility scores $q^*(e_i, e_j)$ and the predicted compatibility scores $\hat{q}(e_i, e_j)$ for all pairs of SFCs (e_i, e_j) in a given training set. If two SFCs e_i are assigned to the same template instance in the gold standard, then $q^*(e_i, e_j) = 1$, otherwise $q^*(e_i, e_j) = 0$, Note that we only consider pairs of slot-filler candidates which are assigned to the same template type.

The complete model is trained by minimizing the loss $L_{EE} + L_{TA}$ with respect to model parameters which are given by the parameters of the BERT encoder, the parameters of the dense layers defined by (1), (2), (8) and the parameters of the layer which is used to compute the vector representation $\mathbf{e}_k$ of the SFCs.

4 Experiments

We conduct experiments on two public datasets (Sanchez-Graillet et al., 2021) which contain RCT abstracts from the Glaucoma and Type 2 Diabetes Mellitus (T2DM) domain, respectively. The corpora of both datasets are annotated at two levels: At the first level, salient entities which describe components of the PICO elements are annotated. The second level comprises template-based annotations of complex PICO elements and their interactions.

4.1 Experimental Setting

In all our experiments, we use a BERT model pretrained on biomedical and life sciences literature abstracts [1]. We use the same train/validation/test split as in (Sanchez-Graillet et al., 2021), Table 2 shows the number of abstracts included in the train, validation and test sets of the respective datasets. All models are trained with the AdamW optimizer (Loshchilov and Hutter, 2017) for 30 epochs with an inital learning rate of $3 * 10^{-5}$ and with a linear warm-up phase over the first 10% of training steps. Further, we use batches of exactly one abstract and set the beam size of the intra-template compatibility optimization algorithm depicted in 1 to 50.

We score a predicted SFC as correct if there is a SFC in the corresponding sentence in the test set with the same label, start and end position. Further, we use the Hungarian algorithm (Kuhn, 1955) for aligning predicted and ground truth templates for

[1] https://tfhub.dev/google/experts/bert/pubmed/2

Table 2: Number of abstracts in the train, validation and test sets

	# Abstracts training set	# Abstracts validation set	# Abstracts test set
Glaucoma	69	17	21
T2DM	68	16	20

each template class, using the pairwise micro F_1 as optimization objective.

As a baseline, we implement a *greedy assignment* approach to assign SFCs to template instances: Given the set $\mathcal{E}_t$ of extracted SFCs for template type t, we repeatedly loop over the template instances T_t^k, randomly pick a SFC from $\mathcal{E}_T$, assign this entity to T_t^k and remove it from $\mathcal{E}_t$. This is repeated until the set $\mathcal{E}_t$ is empty, i.e., all SFCs for template type t have been assigned.

4.2 Results

Extraction of slot filler candidates: Our approach can extract 37 types of slot filler candidates (see Table 1). The results in terms of Precision, Recall and F-Measure for all slot types are given in Table 8 in the Appendix. Overall, the model yields a micro-averaged F-Measure of 0.80 (P=0.80, R=0.73) on the Glaucoma dataset as well as F=0.76 (P=0.80, R=0.73) on the T2DM dataset. Table 3 shows the top 10 slot types with the best extraction results. Similarly, table 4 shows the five slot types with the worst prediction results.

Template extraction: Table 8 in the Appendix shows the prediction results of the SFCs on the Glaucoma and T2DM test sets. The entries "-" indicate that the corresponding slots are not used in the respective data set. Table 5 shows the aggregated results over each template type by averaging the F-values for all slots of the corresponding template. Note that Table 5 only contains template types which could have more than one instance, whereas Table 1 shows all template types. Overall, our proposed model yields a micro F_1 score of 62.27% on the Glaucoma corpus and 64.38% on the T2DM corpus, with a gain of 6,34% in micro-averaged F_1 compared to greedy assignment on the Glaucoma dataset and 3,95% on the T2DM dataset, showing the superiority of our proposed intra-template entity compatibility (ITC) algorithm. For both datasets, the instances of template Arm are extracted best with mean F_1 of 91% and 93% on the Glaucoma and T2DM dataset, respectively. The templates types that have the worst performance are Endpoint for the Glaucoma dataset (mean F=48%)

Table 3: Top 10 slot types for the Glaucoma and T2DM datasets

Slot Name	F_1
Glaucoma	
PMID	1.00
PublicationYear	1.00
RelativeChangeValue	1.00
SdErrorChangeValue	1.00
Title	0.94
SdDevResValue	0.94
NumberPatientsCT	0.93
ChangeValue	0.92
HealthCondition	0.91
NumberPatientsArm	0.91
T2DM	
NumberAffected	1.00
PMID	1.00
PublicationYear	1.00
Journal	0.97
PercentageAffected	0.95
Author	0.94
NumberPatientsArm	0.93
NumberPatientsCT	0.93
ChangeValue	0.90
CTDesign	0.88

Table 4: Slot types with the worst prediction results for the Glaucoma and T2DM datasets

Slot Name	F_1
Glaucoma	
ObservedResult	0.00
Drug	0.27
Precondition	0.28
PointDescription	0.32
ObjectiveDescrip-	0.49
T2DM	
ConfIntervalDiff	0.00
ObservedResult	0.00
SdDevChangeValue	0.25
SdDevBL	0.38
Precondition	0.41

h

Table 5: Aggregated slot-filling results (mean F_1 and overall micro F_1) (ITC=Intra-Template Compatibility)

	Glaucoma		T2DM	
	Greedy Assignment	ITC	Greedy assignment	ITC
DiffBetweenGroups	0.58	**0.64**	0.47	**0.48**
Arm	0.85	**0.91**	0.93	0.93
Intervention	0.53	**0.73**	**0.68**	0.58
Medication	0.89	0.89	0.57	**0.77**
Outcome	0.41	**0.61**	**0.47**	0.44
Endpoint	**0.51**	0.48	0.56	**0.60**
Micro Average	0.56	**0.62**	0.60	**0.64**

and Outcome for the T2DM dataset (mean F=44%).

On the Glaucoma dataset, for four out of six template types, our proposed ITC algorithm yields better performance than the greedy assignment in terms of mean F_1, for one template type (Medication) the performance is equal and for one out of six template types the performance is worse. On the T2DM dataset, for three out of six template types our ITC algorithm performs better than greedy assignment, for one template type (Arm) the performance is equal and and for two out out six template types the performance is worse.

We also conducted a study simulating perfect entity extraction by performing the second step with gold standard SFCs. The results in Table 6 show that results are significantly better with perfect SFC identification, yielding an increase of more than 0.20 points in micro averaged F-Measure for the Glaucoma dataset and more than 0.15 points on the T2DM dataset. This shows the importance of good entity recognition and extraction models.

Table 7 shows the effect of the beam size on the template extraction results. Overall, we see that the beam size has a negligible effect on the results.

Case study: As a case study, we compare the predicted structure to the gold standard structure for one published clinical trial in the test set of the T2DM corpus. We cherry pick the study with the best results in terms of micro-averaged F_1, that is $F_1 = 0.85$. The selected paper is the publication by Shankar et al. (2017). Table 10 contrasts the instances of templates specified in the gold standard vs. the instances of templates extracted by our approach. Overall, the results are very good, clearly showing the potential of our approach and hinting at the fact that the task can be solved to a satisfactory extent. Regarding the *Population* studied in the paper, our method can extract a corresponding condition, but is not able to explicitly extract the countries in which the population was recruited (USA, Australia). With except of the health condition (type 2 diabetes mellitus), all other elements describing the characteristics of the *Clinical Trial* are extracted correctly. Most of the relevant endpoints are extracted correctly, albeit not always the correct units are extracted. Two endpoints are conflated into one: fasting plasma glucose and 2 - h post - meal glucose with the result that one endpoint has a unit (mg/dl) but no endpoint description. The medications for the two arms (sinagliptin vs. placebo) are extracted correctly. The dose value of sinagliptin is mistaken for the dose value of the placebo unfortunately. Most of the outcome values are extracted correctly, but the percentage of patients affected is not extracted. The p values reporting significance of results when comparing the two arms / groups are extracted perfectly.

Table 11 shows the instances of templates specified in the gold standard vs. the instances of templates extracted by our approach for the abstract from the T2DM test set with the worst prediction result in terms of micro $F_1 = 0.57$. The corresponding publication can be found in (Klein et al., 2014). Although our system gets the Publication metadata, the Clinical Trial design, Arms and Medications right to a great extent, it makes a number of important errors in the categories Endpoints and Outcomes.

5 Conclusion

We have presented a twp-step neural architecture based on a transformer model that can induce a structured representation from an abstract describing a Randomized Controlled Trial (RCT). The architecture performs extraction of candidate slot fillers as a first step by identifying spans of 37

Table 6: Aggregated slot-filling results comparing the settings with perfect entity recognition using gold standard entity annotations and entity recognition by our model (mean F_1 and overall micro F_1)

	Glaucoma		T2DM	
	Ground Truth SFCs	Predicted SFCs	Ground Truth SFCs	Predicted SFCs
DiffBetweenGroups	**0.87**	0.64	**0.77**	0.48
Arm	**1.00**	0.91	**1.00**	0.93
Intervention	**0.83**	0.73	**1.00**	0.58
Medication	**0.92**	0.89	**0.93**	0.77
Outcome	**0.69**	0.61	**0.59**	0.44
Endpoint	**0.90**	0.48	**0.82**	0.60
Micro Average	**0.83**	0.62	**0.81**	0.64

Table 7: Effect of the beam size on template extraction results (mean F_1 and overall micro F_1)

	Glaucoma					T2DM				
	10	20	30	40	50	10	20	30	40	50
DiffBetweenGroups	0.64	0.60	0.60	0.60	0.64	0.50	0.48	0.48	0.47	0.48
Arm	0.91	0.91	0.91	0.91	0.91	0.93	0.93	0.93	0.93	0.93
Intervention	0.73	0.73	0.73	0.73	0.72	0.58	0.58	0.58	0.58	0.58
Medication	0.89	0.89	0.89	0.89	0.77	0.77	0.79	0.79	0.77	0.77
Outcome	0.58	0.61	0.57	0.56	0.61	0.42	0.43	0.42	0.43	0.44
Enpoint	0.48	0.48	0.48	0.48	0.48	0.62	0.61	0.60	0.62	0.60
Micro Average	0.62	0.62	0.62	0.62	0.62	0.64	0.64	0.64	0.64	0.64

different classes. At a second step, it assigns the extracted candidate slot fillers into nine main templates. We have shown that our approach can extract candidate slot fillers reliably, yielding micro F-Measures of 76.21% and 76.49% on our Glaucoma and T2DM dataset, respectively. In terms of extraction of templates, our approach yields micro F-measures of 62.27% and 64.38% averaged over all slots on our Glaucoma and T2DM dataset, respectively. The structure of our templates is inspired by the C-TrO ontology (Sanchez-Graillet et al., 2019) and induces the most fine-grained and accurate representation of a published RCT that has been considered so far by any information extraction system. In future work we intend to show that our information extraction approach indeed supports the aggregation of results across clinical trials. Further, we plan to use the intra-template compatibility scores to infer the number of template instance for template types which could have several instances. This can be regarded as an additional layer on top of our proposed optimization algorithm. In addition, we plan to predict links between template instances.

References

Heike Adel, Laura Ana Maria Oberländer, Sean Papay, Sebastian Padó, and Roman Klinger. 2018. Dere: A task and domain-independent slot filling framework for declarative relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 42–47.

Abien Fred Agarap. 2018. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

Elaine M Beller, Joyce Kee-Hsin Chen, Una Li-Hsiang Wang, and Paul P Glasziou. 2013. Are systematic reviews up-to-date at the time of publication? *Systematic Reviews*, 2:36.

Rohit Borah, Andrew W Brown, Patrice L Capers, and Kathryn A Kaiser. 2017. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the prospero registry. *BMJ Open*, 7(2).

Florian Boudin, Jian-Yun Nie, Joan C Bartlett, Roland Grad, Pierre Pluye, and Martin Dawes. 2010. Combining classifiers for robust pico element detection. *BMC medical informatics and decision making*, 10(1):1–6.

Grace Y Chung. 2009. Sentence retrieval for abstracts of randomized controlled trials. *BMC Medical Informatics and Decision Making*, 9:10.

Dina Demner-Fushman and Jimmy Lin. 2007. Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, 33(1):63–103.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Giacomo Frisoni, Gianluca Moro, and Antonella Carbonaro. 2021. A survey on event extraction for natural language understanding: Riding the biomedical literature wave. *IEEE Access*, 9:160721–160757.

Kenji Hirohata, Naoaki Okazaki, Sophia Ananiadou, and Mitsuru Ishizuka. 2008. Identifying sections in scientific abstracts using conditional random fields. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Di Jin and Peter Szolovits. 2018. Pico element detection in medical text via long short-term memory neural networks. In *Proceedings of the BioNLP 2018 workshop*, pages 67–75.

David J Klein, Tadej Battelino, DJ Chatterjee, Lisbeth V Jacobsen, Paula M Hale, Silva Arslanian, and NN2211-1800 Study Group. 2014. Liraglutide's safety, tolerability, pharmacokinetics, and pharmacodynamics in pediatric type 2 diabetes: a randomized, double-blind, placebo-controlled trial. *Diabetes technology & therapeutics*, 16(10):679–687.

GG Koch. 2006. No improvement–still less than half of the cochrane reviews are up to date. In *XIV Cochrane Colloquium, Dublin, Ireland*.

F.R. Kschischang, B.J. Frey, and H.-A. Loeliger. 2001. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519.

Harold W Kuhn. 1955. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Feifan Liu, Jinying Chen, Abhyuday Jagannatha, and Hong Yu. 2016. Learning for biomedical information extraction: Methodological review of recent advances. *arXiv preprint arXiv:1606.07993*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Larry McKnight and Padmini Srinivasan. 2003. Categorization of sentence types in medical abstracts. In *AMIA Annual Symposium Proceedings*, volume 2003, page 440. American Medical Informatics Association.

Benjamin E Nye, Ani Nenkova, Iain J Marshall, and Byron C Wallace. 2020. Trialstreamer: mapping and browsing medical evidence in real-time. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2020, page 63. NIH Public Access.

Annette M O'Connor, Guy Tsafnat, Stephen B Gilbert, Kristina A Thayer, Ian Shemilt, James Thomas, Paul Glasziou, and Mary S Wolfe. 2019. Still moving toward automation of the systematic review process: a summary of discussions at the third meeting of the international collaboration for automation of systematic reviews (icasr). *Systematic reviews*, 8:57.

Yannis Papanikolaou and Francine Bennett. 2021. Slot filling for biomedical information extraction.

David L Sackett, William MC Rosenberg, JA Muir Gray, R Brian Haynes, and W Scott Richardson. 1996. Evidence based medicine. *BMJ*, 313(7050):170.

Olivia Sanchez-Graillet, Philipp Cimiano, Christian Witte, and Basil Ell. 2019. C-TrO: An Ontology for Summarization and Aggregation of the Level of Evidence in Clinical Trials. In *Proc. of the 5th Joint Ontology Workshops (JOWO): Ontologies and Data in the Life Sciences*.

Olivia Sanchez-Graillet, Christian Witte, Frank Grimm, and Philipp Cimiano. 2021. An annotated corpus of clinical trial publications supporting schema-based relational information extraction. *Journal of Biomedical Semantics*. Under review.

R Ravi Shankar, Yuqian Bao, Ping Han, Ji Hu, Jianhua Ma, Yongde Peng, Fan Wu, Lei Xu, Samuel S Engel, and Weiping Jia. 2017. Sitagliptin added to stable insulin therapy with or without metformin in chinese patients with type 2 diabetes. *Journal of diabetes investigation*, 8(3):321–329.

Antonio Trenta, Anthony Hunter, and Sebastian Riedel. 2015. Extraction of evidence tables from abstracts of randomized clinical trials using a maximum entropy classifier and global constraints. *arXiv preprint arXiv:1509.05209*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Xing David Wang, Leon Weber, and Ulf Leser. 2020. Biomedical event extraction as multi-turn question answering. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 88–96.

Stephen Wu, Kirk Roberts, Surabhi Datta, Jingcheng Du, Zongcheng Ji, Yuqi Si, Sarvesh Soni, Qiong Wang, Qiang Wei, Yang Xiang, et al. 2020. Deep learning in clinical natural language processing: a methodical review. *Journal of the American Medical Informatics Association*, 27(3):457–470.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

A Supplementary Material

Table 8: Results of the slot-filler candidate extraction on the Glaucoma and T2DM test sets

	Glaucoma			T2DM		
Slot Name	Precision	Recall	F_1	Precision	Recall	F_1
analysesHealthCondition	0.97	0.86	0.91	0.64	0.58	0.61
Author	1.00	1.00	1.00	0.88	1.00	0.94
BaselineUnit	0.62	0.48	0.54	0.81	0.81	0.81
BaselineValue	0.90	0.67	0.77	0.80	0.60	0.69
CTDesign	0.76	0.83	0.79	0.85	0.91	0.88
CTduration	0.84	0.94	0.89	0.80	0.84	0.82
ChangeValue	0.97	0.88	0.92	0.90	0.90	0.90
ConclusionComment	0.85	0.79	0.81	0.83	0.31	0.45
ConfIntervalDiff	-	-	-	0.00	0.00	0.00
Country	0.81	0.89	0.85	0.89	0.44	0.59
DiffGroupAbsValue	0.75	0.67	0.71	0.84	0.70	0.76
DoseUnit	0.61	0.82	0.70	0.84	0.80	0.82
DoseValue	0.72	0.68	0.70	0.87	0.73	0.80
Drug	0.40	0.21	0.27	0.84	0.76	0.80
EndPointDescription	0.32	0.33	0.32	0.68	0.80	0.74
Frequency	0.89	0.71	0.79	0.71	0.57	0.63
Journal	0.76	0.76	0.76	1.00	0.95	0.97
NumberAffected	0.63	1.00	0.77	1.00	1.00	1.00
NumberPatientsArm	0.88	0.94	0.91	1.00	0.87	0.93
NumberPatientsCT	0.93	0.93	0.93	0.93	0.93	0.93
ObjectiveDescription	0.56	0.43	0.49	0.50	0.44	0.47
ObservedResult	0.00	0.00	0.00	0.00	0.00	0.00
PMID	1.00	1.00	1.00	1.00	1.00	1.00
PValueChangeValue	0.50	0.75	0.60	0.83	0.45	0.59
PercentageAffected	0.82	0.95	0.88	0.96	0.94	0.95
Precondition	0.42	0.22	0.29	0.57	0.32	0.41
PublicationYear	1.00	1.00	1.00	1.00	1.00	1.00
PvalueDiff	0.49	0.68	0.57	0.82	0.94	0.88
RelativeChangeValue	1.00	1.00	1.00	-	-	-
RelativeFreqTime	0.44	0.67	0.53	-	-	-
ResultMeasuredValue	0.85	0.79	0.82	0.75	0.95	0.84
SdDevBL	1.00	0.67	0.80	0.60	0.27	0.38
SdDevChangeValue	0.89	0.67	0.76	0.22	0.29	0.25
SdDevResValue	0.87	1.00	0.93	0.41	1.00	0.58
SdErrorChangeValue	1.00	1.00	1.00	-	-	-
TimePoint	0.60	0.71	0.65	0.63	0.57	0.60
Title	1.00	0.88	0.94	0.77	0.77	0.77
micro average:	0.80	0.73	0.76	0.80	0.73	0.77

Table 9: F_1 scores of the assignment of slot-filler candidates to template instances on the Glaucoma and T2DM test sets. The ITC (Intra-Template Compatibility) columns show the results of our proposed method

	Glaucoma		T2DM	
Slot Name	Greedy Assignment	ITC	Greedy Assignment	ITC
DiffBetweenGroups				
ConfIntervalDiff	-	-	0.00	0.00
PvalueDiff	0.57	0.57	0.83	**0.83**
DiffGroupAbsValue	0.59	**0.71**	0.58	**0.62**
mean	0.58	**0.64**	0.47	**0.48**
Arm				
NumberPatientsArm	0.85	**0.91**	**0.93**	0.92
mean	0.85	**0.91**	0.93	0.93
Intervention				
Frequency	0.79	0.79	**0.68**	0.58
RelativeFreqTime	0.27	**0.67**	-	-
mean	0.53	**0.73**	**0.68**	0.58
Medication				
Drug	**0.37**	0.34	0.73	**0.83**
DoseValue	**0.70**	0.65	0.46	**0.63**
DoseUnit	0.71	**0.79**	0.49	**0.87**
mean	0.89	0.89	0.57	**0.77**
Outcome				
ResultMeasuredValue	**0.44**	0.41	**0.61**	0.56
TimePoint	**0.55**	0.50	**0.55**	0.35
PValueChangeValue	0.40	**0.42**	0.59	0.59
PercentageAffected	0.65	**0.65**	0.72	**0.89**
SdErrorChangeValue	0.29	**1.00**	-	-
BaselineValue	0.39	**0.69**	0.34	**0.46**
SdDevBL	0.56	**0.80**	**0.25**	0.13
RelativeChangeValue	0.00	**1.00**	-	-
ChangeValue	**0.79**	0.73	**0.73**	**0.71**
SdDevResValue	0.37	**0.74**	0.42	0.42
NumberAffected	0.46	0.46	**0.88**	0.50
SdDevChangeValue	0.48	**0.57**	0.13	**0.25**
ObservedResult	0.00	0.00	0.00	0.00
mean	0.41	**0.61**	**0.47**	0.44
Endpoint				
EndPointDescription	**0.29**	0.28	0.56	**0.63**
BaselineUnit	**0.73**	0.68	0.55	**0.57**
mean	**0.51**	0.48	0.60	**0.64**
micro average	0.56	**0.62**	0.60	**0.64**

	Gold Standard	Predicted
	Population	
Country	usa, australia	
Precondition	chinese patients with type 2 diabetes mellitus receiving stable insulin therapy alone or in combination with metformin	patients with inadequate glycemic control on insulin (glycated hemoglobin (hba1c) at 7.5% and at 11%)
	Publication	
Author	shankar rr | bao y | han p | hu j | ma j | peng y | wu f | xu l | engel ss | jia w	engel ss | shankar rr | bao y | han p | hu j | ma j | peng y | wu f | xu l | jia w
Journal	j diabetes investig	j diabetes invest ig .
PMID	27740719	27740719
PublicationYear	2017	2017
Tile	sitagliptin added to stable insulin therapy with or without metformin in chinese patients with type 2 diabetes .	sitagliptin added to stable insulin therapy with or without metformin in chinese patients with type 2 diabetes .
	Clinical Trial	
healthCondition	type 2 diabetes mellitus	
Design	randomized	randomized
Duration	24 weeks	24 weeks
NumberPatients	467	467
ObjectiveDescription	we evaluated the tolerability and efficacy of the addition of sitagliptin in chinese patients with type 2 diabetes mellitus receiving stable insulin therapy alone or in combination with metformin.	we evaluated the tolerability and efficacy of the addition of sitagliptin in chinese patients with type 2 diabetes mellitus receiving stable insulin therapy alone or in combination with metformin .
	Endpoints	
BaselineUnit:	%	
EndPointDescription	hba1c	hba1c
EndPointDescription	hba1c of < 7.0%	hba1c of < 7.0%
BaselineUnit	mg / dl	mg / dl
EndPointDescription	2 - h post - meal glucose	fasting plasma glucose, 2 - h post - meal glucose
BaselineUnit	mg / dl	mg/dl
EndPointDescription	fasting plasma glucose	
BaselineUnit	mg / dl	
EndPointDescription	hypoglycemia (symptomatic or asymptomatic)	hypoglycemia
BaselineUnit:		
EndPointDescription	bodyweight	bodyweight
	Medications	
DoseUnit	mg	mg
DoseValue	100	
Drug	sitagliptin	sitagliptin
DoseUnit	100	
DoseValue		
Drug	placebo	placebo
	Outcomes	
ChangeValue	0.7	
ChangeValue	0.3	0.3
PercentageAffected	16	
TimePoint	week 24	week 24
PercentageAffected	8	
ChangeValue	26.5	26.5
ChangeValue	14.4	14.4
ChangeValue	10.7	10.7
NumberAffected	64	
PercentageAffected	27.4	27.4
NumberAffected	51	
PercentageAffected	21.9	21.9, 8
ObservedResult	neither group had a significant change from baseline in bodyweight.	
	Differences between groups	
PvalueDiff	p < 0.001	p < 0.001
PvalueDiff	p = 0.013	p = 0.013
PvalueDiff	p < 0.001	p < 0.001

Table 10: Predicted and gold standard structures for the abstract of the clinical trial described in Shankar, R Ravi et al. "Sitagliptin added to stable insulin therapy with or without metformin in Chinese patients with type 2 diabetes." Journal of diabetes investigation vol. 8,3 (2017): 321-329. doi:10.1111/jdi.12585; within one template type, horizontal lines separate different instances of the same template type

	Gold Standard	Predicted
	Population	
AvgAge	14 . 8	
Country	ohio	
MaxAge	17	
MinAge	10	
Precondition	youth treated with diet / exercise alone or with metformin and having a hemoglo bin a1c (hba1c) level of 6 . 5 - 11 %; youth	
	Publication	
Author	battelino t; arslanian s; jacobsen lv; chatterjee dj; klein dj; hale pm	lopez x; neufeld n; battelino t; blumer j; arslanian s; bone m; randell t; jacobsen lv; chatterjee dj; hazan l; ferry r; christensen m; tsalikian e; toltzis p; de schepper j; wadwa rp; wintergerst k; klein dj; barrett t; hale pm
Journal	diabetes technol ther .	diabetes technol ther .
PMID	25036533	25036533
PublicationYear	2014	2014
	Clinical Trial	
analysesHealthCondition	type 2 diabetes	type 2 diabetes
CTDesign	randomized | double - blind	randomized
CTduration	5 weeks	5 weeks
	Arms	
NumberPatientsArm	14	14
NumberPatientsArm	7	7
	Endpoints	
EndPointDescription	severe hypoglycemia	hba1c
EndPointDescription	gastrointestinal aes	
EndPointDescription BaselineUnit	hba1c %	 %
EndPointDescription BaselineUnit	body weight kg	 kg
	Medications	
DoseUnit Drug	mg liraglutide	mg placebo | liraglutide
	Outcomes	
ObservedResult hasSdDevBL	no serious adverse events 35 . 6	
ObservedResult TimePoint	were most common at lower liraglutide doses during dose escalation . 5 weeks	
ResultMeasuredValue BaselineValue	12 113 . 2	
ResultMeasuredValue	1 . 7	1 . 7 | 0 . 3
ChangeValue TimePoint	0 . 86 5 weeks	0 . 86
ChangeValue	0 . 04	0 . 50 | 0 . 04
ChangeValue BaselineValue	0 . 50 8 . 1	
ChangeValue	0 . 54	0 . 54
	Differences between groups	
hasPvalueDiff	p = 0 . 9703	p = 0 . 9703
hasPvalueDiff	p = 0 . 0007	p = 0 . 0007

Table 11: Predicted and gold standard structures for the abstract of the clinical trial described in Klein, David J et al. "Liraglutide's safety, tolerability, pharmacokinetics, and pharmacodynamics "pediatric type 2 diabetes: a randomized, double-blind, placebo-controlled trial" Diabetes technology & therapeutics vol. 16,19 (2014): 679-687. doi:10.1089/dia.2013.0366; within one template type, horizontal lines separate different instances of the same template type; "|" separates SFCs