Information Sciences 662 (2024) 120232

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Validating and constructing behavioral models for simulation and projection using automated knowledge extraction

Tabea S. Sonnenschein [a,b,c,*], G. Ardine de Wit [b,d,f], Nicolette R. den Braver [e], Roel C.H. Vermeulen [b,c], Simon Scheider [a]

[a] Human Geography and Spatial Planning, Faculty of Geosciences, Utrecht University, the Netherlands
[b] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands
[c] Institute of Risk Assessment Sciences, Utrecht University, the Netherlands
[d] Centre for Nutrition, Prevention and Healthcare, National Institute of Public Health and the Environment (RIVM), the Netherlands
[e] Department of Epidemiology & Data Science, Amsterdam Public Health Research Institute, Amsterdam University Medical Centres, Amsterdam, the Netherlands
[f] Vrije Universiteit Amsterdam, Faculty of Science, Department of Health Sciences & Amsterdam Public Health Research Institute, Amsterdam, the Netherlands

ARTICLE INFO

Dataset link: https://knowledgesynth.github.io/ontologies/bcdo.html

Keywords:
Validation
Knowledge extraction
Knowledge synthesis
Knowledge graph
Behavior modeling
Simulation
Ontology
BERT
Named-entity recognition

ABSTRACT

Human behavior may be one of the most challenging phenomena to model and validate. This paper proposes a method for automatically extracting and compiling evidence on human behavior determinants into a knowledge graph. The method (1) extracts associations of behavior determinants and choice options in relation to study groups and moderators from published studies using Natural Language Processing and Deep Learning, (2) synthesizes the extracted evidence into a knowledge graph, and (3) sub-selects the model components and relationships that are relevant and robust. The method can be used to either (4a) construct a structurally valid simulation model before proceeding with calibration or (4b) to validate the structure of existing simulation models. To demonstrate the feasibility of the method, we discuss an example implementation with mode of transport as behavior choice. We find that including non-frequently studied significant behavior determinants drastically improves the model's explanatory power in comparison to only including frequently studied variables. The paper serves as a proof-of-concept which can be reused, extended or adapted for various purposes.

1. Introduction

Many societal problems are closely intertwined with human behavior, such as public health, climate change and crime [20,58,52,46]. To comprehend these complex problems and evaluate potential solutions, we need to model human behavior and how it changes in different intervention scenarios. For instance, one may inquire how an enhanced quality of biking infrastructure influences the decision of city residents to opt for bicycling as a means of commuting to work. Agent-based modeling (ABM) is an example of an explicit process-based simulation method that could be used to analyze such behavioral scenarios [9]. In ABM, agents (e.g., urban residents) interact with each other and their environment. These interactions alter the attributes of agents and the environment,

* Corresponding author at: Institute of Risk Assessment Sciences, Utrecht University, the Netherlands.
E-mail address: t.s.sonnenschein@uu.nl (T.S. Sonnenschein).

https://doi.org/10.1016/j.ins.2024.120232
Received 6 June 2023; Received in revised form 22 January 2024; Accepted 24 January 2024

Available online 7 February 2024

which both can be represented using geographic information [16]. A behavioral model in that context involves an agent gathering information about its environment (modeling human perception or communication) and making decisions based on their individual attributes and the environment. One of the most challenging aspects of such a simulation model is the validation of the behavioral component [6,8].

In simulation modeling, model validation means "substantiating that the behavior of the model represents the behavior of the system with sufficient accuracy" [8]. Many modelers concentrate their validation efforts solely on calibrating and cross-validating model parameters. However, this approach is not sufficient to ensure valid estimations as it might not accurately reflect the actual causal structure, potentially missing important confounding variables or interaction effects or including irrelevant variables that contaminate the model. Particularly when aiming to *project* behavioral changes following a hypothetical intervention or scenario [55], the validity of the model structure including the variable selection is crucial for producing trustworthy results that decision-makers can build upon. Yet, there remains a lack of rigorous and efficient methodologies for performing *structural model validation* [8]. In this paper, we propose leveraging existing knowledge of the causal structure of behavior to validate the model structure before calibration.

Utilizing decades of research in the behavioral sciences to generate a behavioral model is a challenging task. The evidence is often scattered, fragmented, occasionally conflicting, and lacks standardized variable names [35,39]. In this context, manually extracting evidence on behavior determinants related to specific types of behavior for specific social groups is extremely labor-intensive. Furthermore, simulation models typically contain multiple modeled behaviors and model components, for which it could be challenging to manually collect and synthesize all knowledge. For this reason, we aim to automate a significant portion of this knowledge extraction process. One major challenge lies in how to represent this knowledge.

Recently, interest in using ontologies to formalize behavioral theories and synthesize empirical evidence has grown [41]. An ontology is a "formal specification of a shared conceptualization" [10], a form of structured knowledge representation using logic. In this sense, "computational ontologies are a means to formally model the structure of a system, i.e., the relevant entities and relations that emerge from its observation, and which are useful to our purposes" [25]. Populating an ontology with behavioral evidence enables us to apply reasoning to make inferences based on the knowledge, to validate the logical consistency of the evidence and to make use of the evidence to validate the structure of a behavioral model. Previous work has focused on understanding the value of ontologies for the behavioral sciences [27,35,39,50], developing domain-specific ontologies [1,40,29,51,15], and enhancing the methodology for behavioral ontology development [32,41–43,60]. Given the abundance of behavioral literature, automating the construction of ontologies of behavioral evidence could help validate the structure of behavioral models more efficiently. There are a few approaches that can automate parts of the processes of ontology construction from text with sufficient quality, such as term extraction or concept hierarchy discovery [3]. However, no method exists yet to automatically extract variables and statistical relations between those variables from scientific literature that together constitute a piece of evidence [3].

In this paper, we propose and test a method for automatically constructing and validating behavioral models for explicit simulations and prediction models based on natural language processing (NLP) and transformers-based deep learning. The method 1) automatically extracts evidence of behavior determinants in relation to choice options, population subgroups and moderators from published studies, 2) synthesizers the extracted evidence in a knowledge graph, 3) sub-selects the model components and relationships that are statistical robust and relevant for the specific model application. We suggest that the method can be used to either 4a) construct a valid model by translating the selected declarative knowledge into procedural code or 4b) validate an existing model by verifying the correspondence of the model to behavioral evidence. To test this validation strategy, we apply our method to the case of modeling urban transport behavior, in particular the choice of the mode of transport. Modal choice directly impacts both health and sustainability. Moreover, it is an ideal type of behavior for applying agent-based simulation as it often involves context-dependent, discrete decisions that require modeling the interaction between a changing environment and changing individual circumstances.

The subsequent section discusses structural model validation as a prerequisite for reliable prediction and simulation of behavior. Section 3 details the methodological steps, followed by the presentation of model performance results, the resulting knowledge graph when applying the method to the literature on modal choice behavior, and a set of example knowledge queries and responses in Section 4. In Section 5, we test the value of the knowledge graph for structural model validation, by showing how it reduces variable selection bias and thereby improves the quality of a transport behavior model. In Section 6, we discuss the benefits and limitations of our method and present directions for future research.

2. Approach to structural model validation

In many studies, the notion of validation is reduced to calibrating parameter settings based on ground truth data and by measuring the fitness of the resulting model on a different data set [16,48]. Aumann [8] calls this *replicative validity*. The validation of the causal model structure, i.e., the representation of variables and relationships, is often neglected. In many cases of predictive behavioral simulations, researchers choose one of the behavioral frameworks that do not inform variable selection[1] and consequently feed it with a set of behavior determinants that can be found in the literature. However, the latter step of variable selection based on literature is most often not done in a systematic way (e.g. [36,4,61,57]), most likely because systematic literature reviews are labor-

[1] (e.g. theory of planned behavior [2], cognitive dissonance theory [21], rational choice theory [38], the beliefs desires intentions model [23] or the PECS framework [47]).

intensive. Some take a qualitative approach by interviewing experts or stakeholders to validate the variable selection [56,54]. While this approach might be the only option when there is a lack of prior evidence and data, it is prone to subjective bias.

However, without a systematic overview of significant behavior determinants and interactions, one could miss important confounding variables, interactions, or moderation effects. In that case, even if the parameters are calibrated with local data, the model would not be able to reliably predict future behavior in which some of the behavioral determinants change in a way that was never observed before, and the model is likely to suffer from low generalizability to other settings. Aumann [8] calls this validation of the model structure, the variable selection, and the types of relationships that are represented, *structural validity*. "Structural validity ensures that the model is generating the correct [input-output] behavior for the right reasons, and not because incorrect behavior of one model component is compensated for by incorrect behaviors of other components." [8]. However, how can one validate a selection of represented variables and relationships?

We know whether a claim is valid (i.e., we come to know the truth) primarily by way of two types of procedures [31]: (1) tests or experiments to justify a claim, (i.e. either by a controlled experiment or observational data), (2) reasoning with pre-existing sources of valid knowledge [30]. In the context of variable selection, the first data-driven approach is paradoxical, since it requires a constrained set of variables for which ground truth data will be collected before applying a variable selection method (e.g. LASSO, ridge, or stepwise regression). It is unclear how one can generate an unbiased set of variables for this purpose. Moreover, it is highly inefficient and challenging if not impossible to collect data on all possible behavior determinants, certainly if some of them turn out to be insignificant. It is therefore more reasonable to assess validity by way of reasoning with existing valid knowledge (i.e. scientific evidence). Yet, there are differences in the methodological rigor that scientific studies use, which influences the quality of the resulting scientific evidence. Robust evidence should be reproducible, address statistical biases and endogeneity, reduce measurement error, perform significance threshold correction for multiple hypothesis testing or a false discovery rate controlling procedure, and take account of causal inference. Due to their methodological inclusion criteria, systematic reviews and meta-analyses may provide a good source of robust evidence.

To validate a structural model based on existing knowledge, one has to answer the following questions for a specific modeled phenomenon and evaluate whether the model represents these relationships:

- What are the possible outcomes of the modeled behavior?
- What are the determinants of the behavior and how does each of them relate to the specific outcomes?
- Are there moderating variables that change the relationship between determinants and behavioral outcomes?
- For which social groups/samples are the different relationships between determinants and behavioral outcomes valid?

Modeling is often a circular process of model construction, calibration, and validation. Our approach rearranges that process for the case of explicit *simulation models* by using an evidence base to simultaneously construct and validate the structure of the model (selection of represented variables and relationships). Subsequently, the parameters can be calibrated and cross-validated based on local data. Ideally, one should test the model on interventional data by calibrating it based on data from before an intervention and testing whether it can predict the observed changes in behaviors.

3. Methods

In this section, we present (1) an ontology design pattern [22] that may be used to build our evidence knowledge graph, (2) a method to automatically extract evidence from scientific articles on the specified behavior choice and populate the ontology, (3) our method application to modal choice behavior, (4) a discussion on methods for using the knowledge graph for model construction and validation and (5) our method for model implementation and evaluation.

3.1. Generic design-pattern of a behavior determinants evidence ontology

Our *Behavior Choice Determinants Ontology (BCDO)*[2] functions as a model for the variables and relationships that determine a specific human choice. The classes and relations of the BCDO are modeled in the web ontology language (OWL)[3] and capture the relationships between behavior choice options, determinants and study groups and moderators as well as the statistical significance of the relationships. An illustration of the ontology design pattern showing its class taxonomy and the object and data properties linking instances of these classes can be found in Fig. 1. We designed the ontology in a bottom-up fashion, by encoding the relevant information of three representative behavioral review studies [28,18,7] into a common pattern. Information is considered relevant when it informs the structure of the behavioral model or allows us to discriminate the quality of the evidence. The design pattern is fairly parsimonious but could be extended if needed.

Class concept definitions as used in Fig. 1 are:

- **Behavior Choice Options** are kinds of activities from which an individual can choose for the specified behavioral choice. They are the values of the *dependent variables* in the explanatory studies from which knowledge is extracted. (e.g. walking, biking for modal choice)

[2] https://knowledgesynth.github.io/ontologies/BehaviorChoiceDeterminantsOntology.ttl.
[3] https://www.w3.org/TR/owl2-primer/.

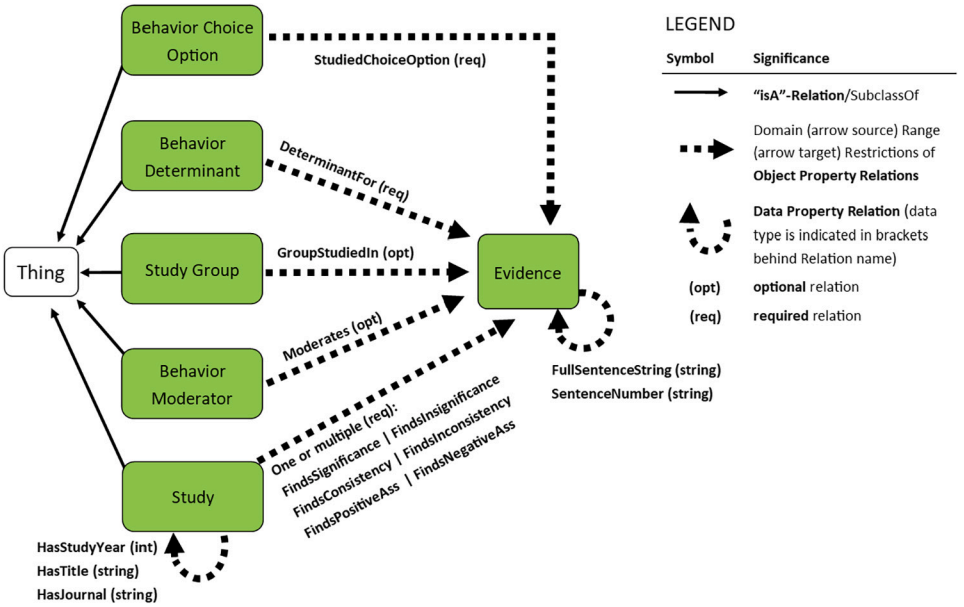


Fig. 1. BCDO Ontology Design Pattern modeled using OWL classes and properties.

- **Behavior Determinants** are factors and constraints that influence the behavioral decision of an individual. These are the determinants or *independent variables* in the evidence provided. (e.g. pedestrian pathway width, gasoline price)
- **Study Group** is the specific *population group or sample* used in the analysis. (e.g. elderly, females)
- **Behavior Moderators** are *variables that moderate* the relationship between a behavior determinant and a behavior choice option. (e.g. household income as a moderator of the importance of gasoline prices for choosing the car)
- **Studies** are scientific articles studying a particular kind of behavior.

Except for the studies themselves, the *instances* (= elements) of the ontology classes are supposed to be *variables* which can be measured. Particular values of the variables are represented in the study data. The variables instantiated in our ontology are by definition measurable because they were extracted from quantitative studies. In the simulation model, corresponding variables will store values that are modified during simulation, for example, infrastructural variables that will be changed during a hypothetical infrastructure intervention or health variables that will be changing depending on air pollution exposure.

In particular, we used the following property relations: *StudiedChoiceOption* links the specific studied choice option to the evidence instance, while *DeterminantFor* does the same for the behavior determinant studied in the evidence instance. *GroupStudiedIn* captures the information on whether the evidence about the statistical association is restricted to a specific social group (i.e. the study group). *Moderates* means that the moderator variable moderates the relation between the behavior determinant and behavior choice option that is part of the same evidence instance. Finally, the statistical association qualifiers have been separated into multiple property relations. Pairs of variable and evidence instances can be connected by more than one relation. *FindsSignificance*, means that a study has found a significant effect of a behavioral determinant, while *FindsInsignificance* indicates the opposite. *FindsConsistency* is a sub-property of *FindsSignificance* and means that a study has found a consistent significant effect of a behavioral determinant across studies. *FindsInconsistency* indicates the opposite and is not a sub-property of *FindsSignificance*. *FindsPositiveAss*, means that a study has found a positive significant effect, while *FindsNegativeAss* indicates that a study has found a negative significant effect. Opposite property relations have been formalized using inverse axioms. This means they are mutually exclusive but not jointly exhaustive.

3.2. Automated extraction and synthesis of prior evidence to populate the BCDO

3.2.1. Justification for method selection

We employ a combination of natural language processing (NLP), bibliometrics, and deep learning to automatically and systematically extract knowledge from previous studies and populate the BCDO. As depicted in Fig. 1, the information that we aim to extract from articles consists of evidence instances, which are composed of variables and qualifiers that describe their statistical relation. In other words, our first main task is to extract the phrases of the articles that correspond to the type of information encoded in our BCDO design pattern. The second main task is to accurately identify the relationships between phrases to form an evidence instance.

Addressing the first task, manual annotation of the entire text data is very time-consuming and hence not scalable. Therefore, we opt for a more efficient automatic approach: training a deep learning model with a small set of labeled documents that will

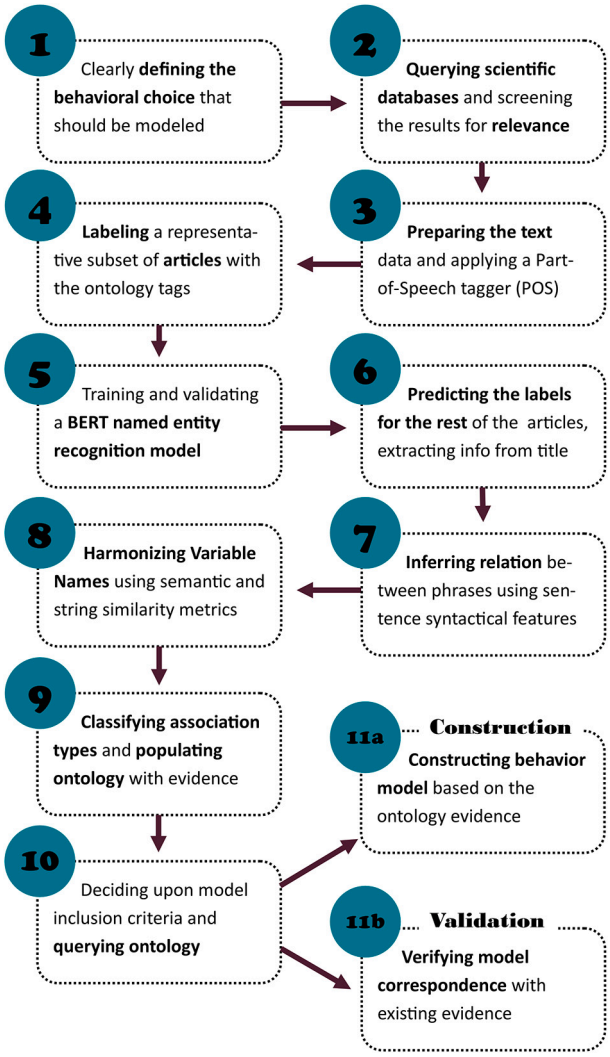


Fig. 2. Knowledge Extraction and Synthesis for Behavior Model Construction and Validation.

subsequently label the rest of the text automatically. For this purpose, we utilize BERT, (Bidirectional Encoder Representations from Transformers) [17], a well-established pretrained NLP model. BERT is trained to understand language context using a masked language model and a next sentence prediction model. It is readily available online in a pretrained format for English (and many other languages) and only requires fine-tuning to a particular NLP task. To classify words with labels, we need to fine-tune BERT to our Named Entity Recognition (NER) task. NER seeks to classify words in a text into predefined categories or tags, which in our case is the assignment of the BCDO tags.

The second task was inferring the relationships between the variables, i.e. determining which phrases form an evidence instance with which other phrases within the sentence. Evidence relations between labeled phrases are expressed with syntactic regularity and can thus be inferred using syntactic features of the sentences. We generated a set of syntactical and keyword-based features of all possible combinations of labeled phrases within the sentences and then trained a machine learning model to predict whether this phrase combination applies or not. In this section, we detail each step as indicated in Fig. 2. We have grouped the steps into: (1) Data Acquisition and Preparation, (2) Named Entity Recognition using BERT, and (3) Evidence Instance Extraction and Synthesis. The Python scripts to perform all of the method steps from step 2 (downloading and preparing the article text data) to step 9 (populating the ontology with the extracted evidence instances), along with complete documentation, can be found on GitHub.[4] From here on we will use quotation marks to indicate extracted text phrases.

[4] https://github.com/KnowledgeSynth/NLP-Knowledge-Extraction-and-Synthesis.

Table 1
Sentence Labeling Examples.

EXAMPLE 1				EXAMPLE 2			
Sen.	Word	POS	BCDO Tag	Sen.	Word	POS	BCDO Tag
1870	Two	CD	O	1009	For	IN	O
1870	articles	NNS	O	1009	example	NN	O
1870	from	IN	O	1009	,	,	O
1870	the	DT	O	1009	most	JJS	I-assocType
1870	same	JJ	O	1009	studies	NNS	O
1870	Belgian	NNP	O	1009	have	VBP	O
1870	study	NN	O	1009	reported	VBN	O
1870	estimated	VBD	O	1009	that	IN	O
1870	the	DT	O	1009	a	DT	O
1870	moderating	VBG	I-assocType	1009	positive	JJ	I-assocType
1870	effect	NN	I-assocType	1009	association	NN	I-assocType
1870	of	IN	O	1009	exists	VBZ	O
1870	area	NN	I-moderator	1009	between	IN	O
1870	level	JJ	I-moderator	1009	well	RB	I-behavDeterm
1870	income	NN	I-moderator	1009	connected	VBN	I-behavDeterm
1870	on	IN	O	1009	street	NN	I-behavDeterm
1870	walkability	NN	I-behavDeterm	1009	networks	NNS	I-behavDeterm
1870	and	CC	O	1009	and	CC	O
1870	total	JJ	I-behavOption	1009	Children's	NNP	I-studygroup
1870	walking	NN	I-behavOption	1009	independent	JJ	I-behavOption
1870	for	IN	I-behavOption	1009	mobility	JJ	I-behavOption
1870	transport	NN	I-behavOption	1009	.	.	O
1870	and	CC	O				
1870	found	VBD	O				
1870	none	NN	I-assocType				
1870	.	.	O				

1. Sen. = Sentence ID; POS = Part of Speech Tag; Citation references have been removed from sentences.
2. Example 1 stems from Cerin et al. [12] and Example 2 from Sharmin & Kamruzzaman [49]

3.2.2. Data acquisition and preparation

Our process begins by defining the specific behavioral choice we aim to model, whether it be modal choice, diet selection, or any other behavior choice (**STEP 1**). The next step involves finding the relevant articles (**STEP 2**) by querying existing scientific databases (e.g. scopus, web of science) with terms related to the behavior choice and screening the results for relevance. To restrict evidence to scientific studies of quality, we recommend selecting only systematic reviews and meta-analyses when possible. Moreover, one can add other quality criteria such as a minimum publication year or relevance criteria that depend on the model goal, for example, a specific studied geographic region or population group. Next, in **STEP 3** one needs to download and prepare the text data of the relevant articles. The output data format should be a data frame of the sequence of words of the text, nested in unique sentence IDs. Additionally, part-of-speech (POS) tags have to be added. POS classify words into categories of words or lexical items that have similar grammatical properties [14]. Examples of such categories are noun plural (NNS), determiner (DT), verb past tense (VBD), verb present tense with 3rd person singular (VBZ), preposition (IN), ect. Downloading the text of the articles has varying levels of difficulty depending on whether it is open access and what format the text is available in. Certain publishers provide full-text XML data of articles, while other articles are only available in PDF format. Our scripts first download XML and PDF documents from a list of DOIs using crossref[5] and Elsevier API.[6] Consequently, they extract and process the text from the documents and apply the POS tagger of python's nltk package [37].

[5] https://pypi.org/project/crossrefapi/1.0.3/. https://www.crossref.org/about/
[6] https://dev.elsevier.com/.

3.2.3. Named-entity recognition using BERT

Subsequently, in **STEP 4** one needs to annotate the words of a subset of scientific articles with the concepts from the BCDO. These annotations will be used to train the BERT model. The labels for annotation include: behavior choice option, behavior determinant, study group, moderator, and association type. Additional data related to the studies themselves (title, journal, year, and DOI, which is the study instance name) can be taken from the bibliometric data provided by the scientific database. Given that a concept may span multiple words, a text chunking strategy is required. We opt for IO-labeling (inside-outside), because it is simple, maximizes the sample size of each tag, reduces computational load, and can capture multi-word phrases [5]. Labels are prefixed with *I-* and all words that do not suit our labels are tagged with *O*. See Table 1 for an example application of IO tagging and Appendix A.1. for our labeling manual. Upon creating the manually labeled training and validation data, the BERT named entity recognition model is trained with it (**STEP 5**). BERT utilizes not only the words and sentence structure but also the POS tags for classifying words into our labels. Concluding the NER task, the unlabeled remaining article's text data processed through the model to predict its labels (**STEP 6**).

The output is the text of the articles, in which each word is listed with its corresponding label (see Table 1). This can be rearranged into a table of sentences with lists of phrases for each label. One can then extract only those sentences that have at least one association-type phrase and at least one behavior determinant, which is the minimum requirement for evidence instances. The other minimum requirement is the behavior choice option to which the behavior determinant has a statistical relation, but these are not necessarily mentioned in the same sentence. Missing behavior choice options can be extracted from the title. Articles focusing on a single behavior choice option or a specific study group typically specify this in the title or abstract. We leverage this pattern to fill in some of the empty study group and behavior choice option slots. Specifically, we check whether any labeled study group or behavior choice option of the article appears in the title and if so, which is the longest string of the respective label that appears in the title. The resulting all-article study group or behavior choice option is consequently assigned to all empty phrase slots of sentences containing evidence instances in the specific article.

3.2.4. Evidence instance extraction and synthesis

After having labeled and processed all the text data, we proceed to **STEP 7**, where we generate evidence instances by establishing the relations between the labeled phrases. These evidence instances will later be used to populate the open slots (classes and property instantiations) of the BCDO. The objective is to construct a table of evidence instances that affirm or refute the significance of a relationship between a behavior choice option, behavior determinant, and optionally, a study group or a moderator. The table also retains the source sentence from which this evidence instance was extracted. To establish the interdependency or relation between phrases within a sentence, we use the syntactical properties of the sentence. For this purpose, we create a table of all possible combinations of association types, behavior options, behavior determinants, study groups and moderators within each sentence. This captures all possible evidence instances of which only a small set actually occurs. We then generate syntactical features at various levels: features pertaining to entire sentences, attributes of elements of the specific instance combinations, and relational attributes of the specific instance combinations. The complete set of 71 features used in this study, along with their construction method, is detailed in Appendix A.2. These features can be used to estimate the veracity of the possible evidence instances using a supervised classification model, such as gradient boosting, support vector machines, a multi-layer perceptron neural net or random forest. The framework allows for the incorporation of additional syntactic features in the future.

Next, in **STEP 8** the variable names need to be harmonized, which is a critical step in the knowledge synthesis. We implement a preliminary approach that utilizes semantic and string similarity metrics and synonym/antonym analysis. To compute string similarity, we decompose variable names into sub-words and evaluate how many sub-words are shared or contained. Moreover, we use the string similarity metric Jaro-Winkler-similarity [59] for harmonizing small spelling variations. For the synonym/antonym identification we use the Python-based nltk wordnet package. Wordnet is an open-source network of semantic relations between words [19]. We start by analyzing how many shared sub-words the different variable names have. If a multi-word variable name is fully contained by another variable, we classify them as synonymous. If one or two sub-words are not contained in the other variable name, but synonyms of these sub-words are present, then they are also classified as synonymous. For single-word variable names contained in another variable name, we filter for variables that are contained in less than 5 other variable names, so that more generic terms (e.g. "density" or "diversity") do not cause noise by linking non-synonymous variables. Moreover, we filter for candidate synonymous variables that are maximum two words long, assuming that any additional words would exceed the possible expressiveness of single-word variable names. For single-word variables, we also check for direct single-word synonyms among the other variable names. Finally, we cast variable name pairs with a Jaro-Winkler similarity above 90 as synonyms.

After generating the synonymity matrix for all variable names relating to all other variable names, we group the variable names into synonymity clusters. All variable names that are synonymous with at least one of the variables of a cluster are joined together. Thereby we assume the transitory property of synonymity. We then identify the optimal variable names for the harmonized variable name clusters by calculating a composite metric of relevance for each variable name within a cluster. This metric is calculated by multiplying (1) the number of directly synonymous variable names to the respective variable name with (2) its number of sub-words and adding (3) its frequency of mentions across scientific articles.[7] We then check if there is a frequently mentioned (four or more mentions) variable name that sticks out, which we choose as the cluster variable name. If there are multiple variable names with

[7] The number of directly synonymous variables decreases with the number of sub-words, which is why we counter that bias by multiplying it by the sub-word length.

the maximum frequency of mentioning in the cluster, we pick the one with the highest composite metric. If there are no frequently mentioned variable names (all below four mentions), we select the one with the highest composite metric. If there are multiple variable names with the highest composite metric value, we choose among them the variable name with the shortest string length. We also incorporate a manual check of the variable harmonization before joining new variable names to the evidence table.

Finally, in **STEP 9** the association type information has to be classified into significant versus insignificant, consistent versus inconsistent, and positive versus negative. Given that academic language used to describe statistical associations is fairly standardized, a keyword-based approach is sufficient and performs well. The resulting relational data of behavior choice option variables, behavior determinant variables, and statistical association qualifiers (as well as study groups and moderators in some cases) can be used to populate the ontology using, for example, the python packages *owlready2*[8] or *rdflib*.[9]

3.3. Application to mode of transport choice

For identifying relevant meta-analyses or systematic review articles, we have used a web of science query using keywords related to transport choice (see Appendix A.3.). We received 582 results, which after screening resulted in 34 relevant papers. We excluded articles in the screening phase if they did not study modal choice determinants or if they did not report on a meta-analysis or systematic review. From these articles, we picked five representative studies, one systematic review, two meta-analyses, and two studies that did both [28,18,7,12,49], which we labeled manually using our BCDO labels: behavior choice option (I-behavOption), behavior determinant (I-behavDeterm), moderator (I-moderator) and study group (I-studygroup). To capture the statistical association, we use (I-assocType). Two labeled example sentences can be found in Table 1.

We subsequently trained our BERT NER model using the labeled data. For this purpose, we used the BERT pretrained cased base model for token classification from Python's transformers package[10] and trained the model using 100 epochs with a batch size of 32 and a learning rate of 0.00003. Since the cased BERT model performed better than the uncased one, it was chosen as the pre-trained model. The five articles that were manually labeled contained 52444 words and 2350 sentences. 20% of the labeled data (i.e. 470 sentences) was used for validation. We used cross-entropy loss to calculate loss and evaluated the model using the F1 score, an evaluation metric to assess the performance of the classification model [24].

For the evidence relation inference, we used the 71 syntactical features listed in Appendix A.2. We manually labeled 4000 of the 151969 possible evidence instances (combinations of phrases) and fed it into multiple supervised classification models (gradient boosting, support vector machines, and random forest) of which the latter performed best. To assess the quality of the evidence relation random forest model, we used repeated stratified 10-fold cross-validation with 3 repeats.

3.4. Using the BCDO for structural model validation

Once the BCDO is populated with evidence instances and hence an evidence base is available, one can make use of it for model construction and validation. Most likely, the knowledge graph contains more factors than one can represent. Hence, **STEP 10** requires querying the knowledge graph for evidence relevant to the specific model one wants to inform. For this purpose, one needs to define criteria for evidence quality and relevance, which can be more or less exclusive. For example, as a quality criterion, one can set a minimum number of studies that have proven the significance of a variable and require that there are no studies with conflicting evidence. To subselect evidence of relevance, one can exclude components if the behavior choice options that one wants to represent are not related to the exogenous variable one is interested in. For example, one may want to only model factors that can be influenced by a specific intervention. Or one may only be interested in a subset of behavior choice options due to their specific relation to, e.g. health.

Once the relevant ontology components are sub-selected, one can either (**STEP 11a**) validate an existing behavioral model by mapping it to the evidence in the BCDO knowledge graph or (**STEP 11b**) construct the behavioral model from scratch using the evidence on significant variables and relationships. For **STEP 11a** one can use queries in the BCDO to determine whether required concepts and object properties are represented. For example, for each modeled behavior choice option, one can query the significant behavior determinants, moderators, or study group heterogeneities and assess whether they are represented in the model. If a deficiency is identified, the model can be improved by including the new variable, interaction, moderation, or heterogeneity. If one is unable to represent factors due to a lack of data, then one can either explicitly acknowledge this as a limitation or use a proxy variable. Moreover, it is possible to perform a sensitivity analysis to quantify the uncertainty introduced when excluding a variable (e.g. [53]).

For **STEP 11b**, more research is needed to identify the best ways to translate the declarative knowledge into procedural code. An agent-based simulation model requires a decision algorithm for, agents [16]. There are a variety of model types, from logic-based architectures, such as beliefs-desires-intentions [38], to ones based on production rules, such as if-then conditions, to probabilistic models, such as probabilistic graphic models [33], to linear multi-criteria decision models, such as utility maximization frameworks. Any of these model types could be fed with the significant behavior determinants and interactions identified in the BCDO knowledge graph.

[8] https://owlready2.readthedocs.io/en/v0.37/.
[9] https://rdflib.readthedocs.io/en/stable/.
[10] https://huggingface.co/docs/transformers/index.

To realize this step, first, one should consider the variable's measurement level and type when deciding on how to represent the variable. For example, one could classify behavior determinants into behavior constraints and behavior factors. Behavior constraints are a boolean type of behavior determinant that limit the choice of an activity and can be represented using if-then conditions. On the other hand, behavior factors are attributes of an individual, their situation, the environment, or the choice option, which make a specific behavior more or less attractive and thus more or less likely to be chosen. Behavior factors (continuous variables) can be transformed into Boolean behavior constraints, but not the other way around.

Secondly, the order of importance of the variables needs to be determined. For predicate logic or production-rule-based models, one can collect data for the significant variables and make use of statistical methods to identify a suitable order of variables and interaction effects (e.g. Bayesian kernel machine regression, importance ranking using decision trees). For Probabilistic Graphic Models or decision tree models, one can use the data to train the model itself. On the other hand, for multi-criteria decision models, one can use the data to calibrate weights for the different variables. Evidence about the heterogeneity of statistical importance of specific behavior determinants between social groups can be represented by using different determinant weights for different social groups. For example, evidence shows that street safety is a more important behavior determinant for women than for men. These heterogeneous weights can also be calibrated using data of the specific social group.

3.5. Implementing and testing the modal choice model based on extracted evidence

To test the impact of the extracted knowledge on behavior prediction, we trained a mode of transport model based on the extracted evidence. We assessed its performance in predicting mode choice using all versus only the most frequently studied variables as suggested by extracted evidence. The model implementation depends on the purpose of the model. For our example, we were interested in an explicit behavior model of the 4 most common modes of transport - biking, walking, driving a car, and public transport use - for scenario simulations of transport interventions in the city of Amsterdam. Hence, we selected evidences that relate to these behavior choice options. We used those evidences where the combination of behavior determinant, studygroup, behavior choice option and a moderator has proven significant more often than it has been proven insignificant. We did not use information on the direction of the statistical association, because that information is lost in the variable harmonization step since variables that might have been encoded differently were grouped into a single variable name. We included studygroups as variables because they can indirectly capture other moderating variables. Then we used the list of unique behavior determinants, studygroups, and moderator variables from our evidence selection to collect data for training a transport behavior model. To train and evaluate this model, we used the ODIN microdata [11] provided in a secured data analysis environment by Statistics Netherlands. ODIN is a representative transport survey from the Netherlands with 179000 trips, of which 6043 trips are made with both origin and destination in Amsterdam, Diemen or Ouder-Amstel. We include the latter two municipalities in the study extent because they have to be traveled through when traveling from the Amsterdam center to Amsterdam Zuidoost. ODIN data includes individual attributes as well as trip origin and destination information on street block level (Postcode 6), which can be used to estimate spatial tracks and to link environmental behavior determinants.

In the extracted evidence we find 123 unique variables that have been identified as linking significantly to at least one of the four behavior choice options of interest. These variables mostly capture attributes related to the built environment, transport network, individual, their social network and social environment. From these variables, we found data to implement 76 of them. We did not find data for psychosocial variables, for aesthetics, littering and vandalism variables as well as for very specific amenities, like easy access building entrances, benches, and cyclist showers in the workspace. We excluded variables that regard "interventions" as a behavior determinant since we are interested in a status quo model that can be used for scenario analysis. An overview of the extracted variables and our measurements of them can be found in Appendix A.4. We linked the environmental variables to the trip survey data, by averaging the raster values along the Euclidean line between the origin and destination. For some variables, we joined the environmental data to only the origin or destination location of the trip (e.g. public transport access, distance to the central business district).

Regarding the implementation of the behavior model, we chose to train decision trees, since they capture various variable interactions well and can be easily interpreted. Since some behavior determinants are correlated, we tested a principal component analysis (PCA) approach on top of using the original set of variables. For this purpose, we calibrated a correlation threshold, such that variables whose correlation with at least one other variable is above the threshold are included in the PCA. The optimal correlation threshold was 0.7. We subsequently included all variables that do not fall above the threshold, plus the resulting principal components as features of the decision tree model. We used 10-fold cross-validation and calibrated all decision tree hyperparameters using grid search. The optimal tree depth was 8, the minimum bucket size 16, and the minimum split size 20. See Appendix A.5 for more information on the hyperparameter and correlation threshold calibration.

To evaluate the added value of the knowledge extraction, we grouped the significant variables into 3 groups: Frequently studied variables (3 or more meta-analyses, now referred to as *FreqVars*), semi-frequently studied variables (2 meta-analyses, now referred to as *SemFreqVars*) and sparsely studied variables (maximum 1 meta-analysis, now referred to as *SparsVars*). We analyzed the additional variance that is captured in the sparsely studied variables by analyzing how much the *SparsVars* correlate with the *SemFreqVars* and *FreqVars* using a correlation matrix. We then tested the model performance (weighted F1 score in predicting validation sample), when including 1) only the *FreqVars*, 2) only *FreqVars* and *SemFreqVars*, and 3) all variables (including the *SparsVars*). Finally, we compared the ranking of variables according to their study frequency with the ranking of variable importance measured by the decision tree method.

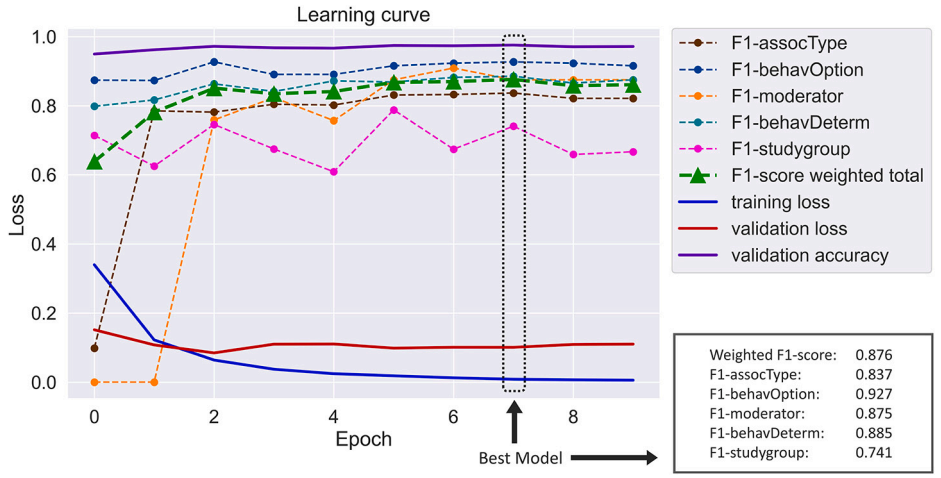


Fig. 3. BERT Learning Curve.

4. Results

In this section, we discuss and evaluate intermediary and final results of our knowledge extraction method when applied to mode of transport choice.

4.1. Knowledge extraction model performance

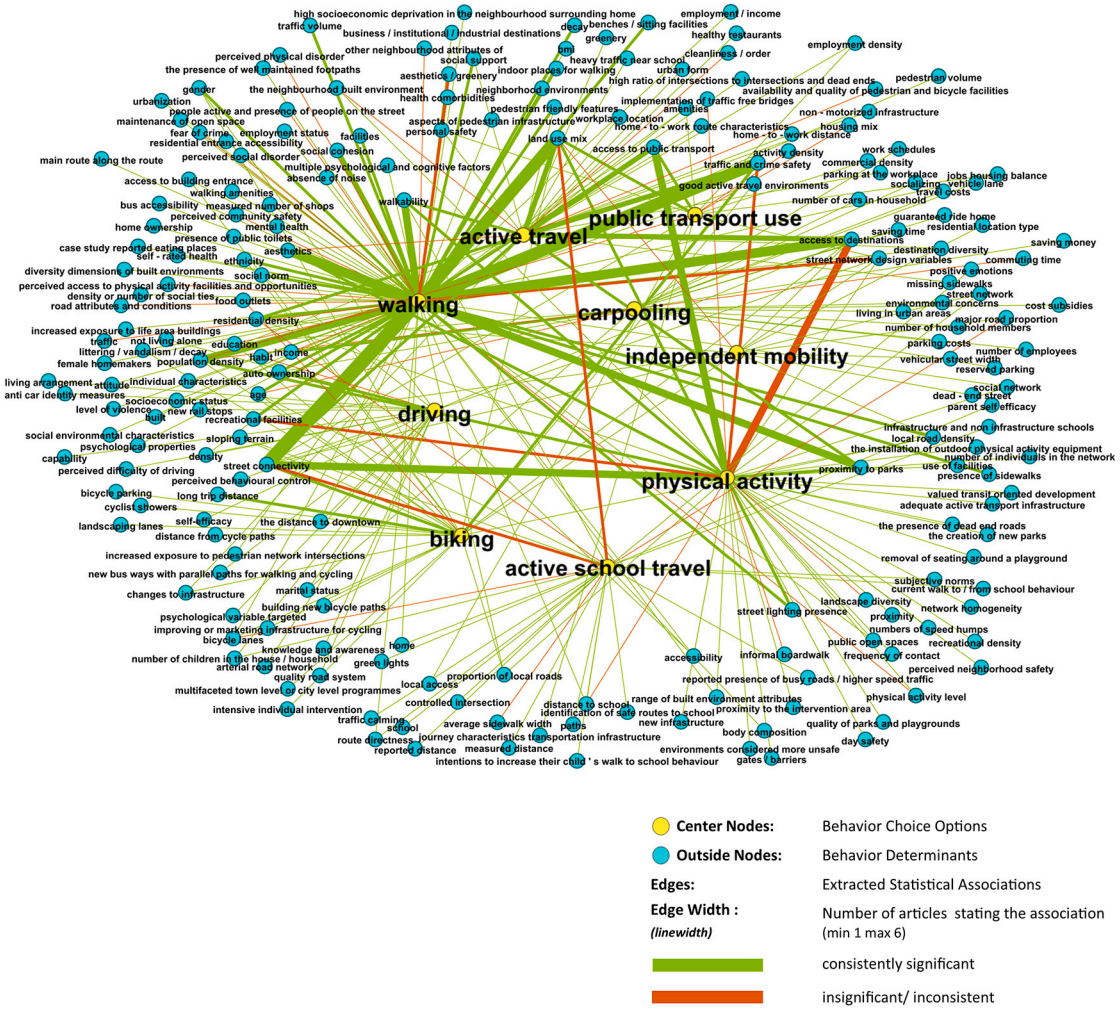
For the labeling of the training sample, we tested inter-annotator agreement by having three researchers annotate a single article using the labeling manual (Appendix A.1). The result was a Cohen's Kappa of 0.88 and 0.93% overlap, indicating substantial agreement. This suggests that the labeling manual provided clear instructions and was effectively used by the researchers. The model performance (F1-score, training loss, validation loss, and validation accuracy) of the BERT NER model across epochs can be found in Fig. 3. The best weighted F1 score was 0.88, indicating a good balance of precision and recall. The random forest model for extracting evidence relations in an article also performed well. The F1 score for predicting true evidence relations was 0.82 and the weighted F1 score was 0.96 (since most of the possible evidence instances are false). It is interesting to note that for both sub-models of our extraction method, namely NER and syntactical feature-based relation classification, precision outperformed recall, likely due to these models' propensity to overfit. This is not necessarily a bad thing, especially because correctness and reliability are arguably more crucial as one can rely on the results obtained from this method. Completeness is harder to achieve, but also less relevant for our task. Despite that, future research should focus on improving recall without compromising precision.

4.2. Results of knowledge synthesis

Through the process of knowledge extraction, we compiled a table encompassing studies, behavior determinants, behavior choice options, reported study groups, and discovered association types. Prior to implementing our variable harmonization method, we identified 395 unique behavior determinants, 91 unique behavior choice options, 27 unique study groups, and 10 unique moderator variables. These variable combinations manifested in 652 unique evidence instances. The scarcity of moderator and study group variables suggests that these aspects receive less focus or reporting in studies. This could also imply that a larger sample of articles might have been beneficial for training due to the relatively small amount of label occurrences in the training dataset.

The automated harmonization process reduced the 91 behavior choice options variable names to 44 categories. Each cluster was manually verified and deemed valid. We further manually aggregated the variable names into 9 categories: walking, active school travel, biking, physical activity, carpooling, independent mobility, active travel, driving, and public transport use. The largest group of variable names, the 395 Behavior determinants, was reduced to 259 variable names by our automated method. This included 57 clusters (synonymous variables) and 202 variable names without identified synonyms. Only two variables have been wrongly assigned to a cluster ("Employment" to "employment density" and "Employment status" to "employment density"). Two pairs of the 57 variable clusters could be fused. One pair that could be fused is related to "public transport access" because *Wordnet*[11] (a synonymity network) does not render "transit" as a synonym of "transport" or vice versa. The other one is "street intersection density" and "street connectivity", which are computed in the same way, but are not composed of synonymous words. Further, 45 of the 202 single variables could have been assigned to a cluster or matched together. After a manual revision, we ended up with 212 unique variable names. See Appendix A.6. for an overview of the variable harmonization of behavior determinants after automatic

[11] https://wordnet.princeton.edu/.



* See https://knowledgesynth.github.io/HarmonisedChoiceOptionsGraph/index.html for an interactive graph visualization

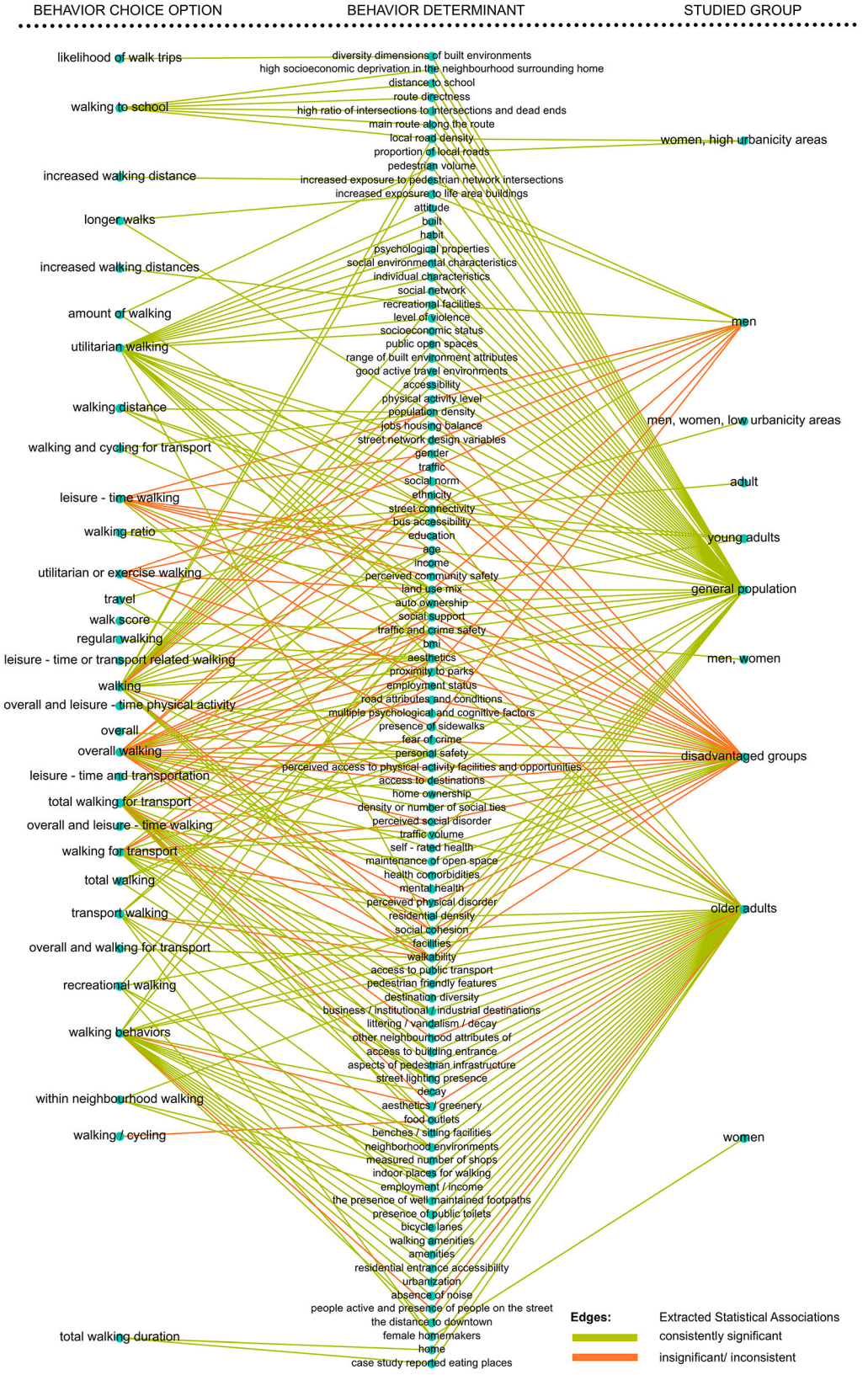
Fig. 4. Simplified Graph of Modal Choice Evidence.

grouping and manual revision. Subsequently, we harmonized the studygroup variable names. The automatic algorithm condensed the 27 variable names to 14, where one cluster was incorrectly formed (men and women, as well as males and females were joined because one was contained in the other). We ended up with 18 variable names after the manual revision. There were a few cases where evidence of synonymous variable names from the same study had been listed twice before variable harmonization. As a result, the variable harmonization step reduced the number of unique evidence instances to 623.

4.3. Resulting knowledge graph

Fig. 4 depicts a simplified version of the resulting knowledge graph. The image does not show the data on study groups, moderators, and study details. The edge width shows the number of articles that have stated the specific association between the harmonized behavior option (center nodes) and behavior determinant (outside nodes). Green edges represent significant and consistent associations across articles while orange means the association is insignificant or inconsistent across articles. It can be seen that walking has been studied most and has the most associated evidence instances (250), followed by physical activity (106) and active travel (56). Moreover, only a few variables have been studied by multiple articles, as can be seen by the small amount of thick and large amount of thin edges. The five most frequently studied variables for modal choice are in that order: land use mix, street connectivity, access to destinations, walkability, and aesthetics. Additionally, a publication bias of significant results can be observed.

Zooming in, Fig. 5 shows the behavior choice options related to *walking* and their statistical associations to behavior determinants and study groups. We depict the harmonized variable names of the behavior determinants and studied groups. However, the behavior choice option variables are the original verbatim variable names as found in the articles. This is to illustrate the challenge of variable harmonization which we address and to show that there is another layer of nuance that distinguishes behavior choice options apart from the 9 grouped modes of transport. One can see that certain study groups have been more predominantly studied than others.



* See https://knowledgesynth.github.io/WalkingGraph/index.html for an interactive graph visualization

Fig. 5. Sub-graph of walking-related Behavior Evidence.

(a) Evidence Instance Example (b) Study Instance Example

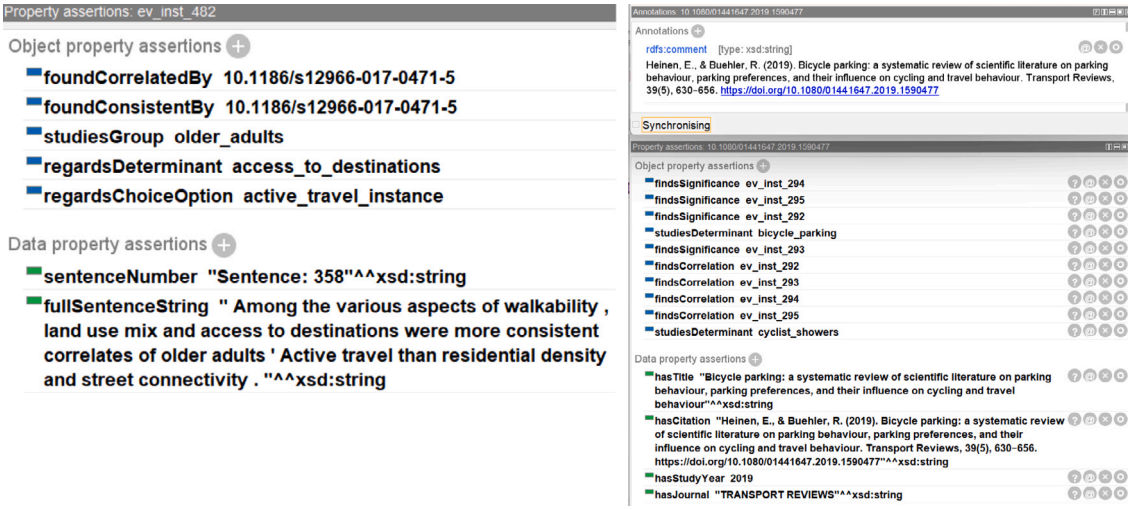


Fig. 6. Ontology Excerpts.

Strikingly, studies on the determinants of walking have often focused on older adults and disadvantaged groups. As far as our extracted knowledge graph shows, men have been studied more than women. See Appendix A.7. for equivalent knowledge graphs for other behavior choices (biking, driving, public transport, and active travel).

4.4. Knowledge queries

Knowledge graph visualizations are only comprehensible to a limited extent and do not capture all of the extracted knowledge. We have populated a knowledge graph with the modal choice evidence using the BCDO, which overcomes these drawbacks and can be queried and processed. The populated graph (and the upper-level ontology) can be downloaded here.[12] It presents the evidence and study information comprehensively (see Fig. 6).

The knowledge graph can be consequently used to query relevant information for model validation and construction. For example, "Find all significant behavior determinants for biking?" can be expressed in SPARQL with:

```
SELECT DISTINCT ?determinant
WHERE {
    ?a a bmo:biking; bmo:studiedChoiceOption ?evidence.
    ?evidence bmo:foundSignificantBy ?study.
    ?evidence bmo:regardsDeterminant ?determinant.
}
```

As a response, we obtain 24 behavior determinants of the extracted evidence that are significantly linked to biking behavior options, such as "bicycle parking", "sloping terrain" and "street connectivity".

We can also formulate more complex questions, such as "What are the 5 most relevant behavioral determinants, measured in terms of how many classes of behavioral choice options they significantly influence?". The formatted response is: "street connectivity" influences 7: independent mobility, public transport use, active travel, biking, driving, physical activity, walking; "access to destinations" influences 6: independent mobility, active travel, biking, driving, physical activity, walking'; "land use mix" influences 6: independent mobility, public transport use, active travel, biking, physical activity, walking; "population density" influences 5: public transport use, active travel, biking, physical activity, walking; "residential density" influences 4: independent mobility, active travel, physical activity, walking. More examples of SPARQL queries and results can be found in Appendix A.8.

5. Evaluation: impact of the extracted evidence on behavior model performance

We plot all three variable groups (defined by their frequency of study) together in a correlation matrix to analyze whether the sparsely studied variables capture new variance or if they are instead highly correlated with frequently studied variables. As Fig. 7 shows, the sparsely studied variables largely do not correlate with the frequently and medium frequently studied variables. The percentage of non-western residents correlates with the distance to the central business district (CBD) (coefficient of 0.58/0.59),

[12] https://knowledgesynth.github.io/ontologies/bcdo.html.

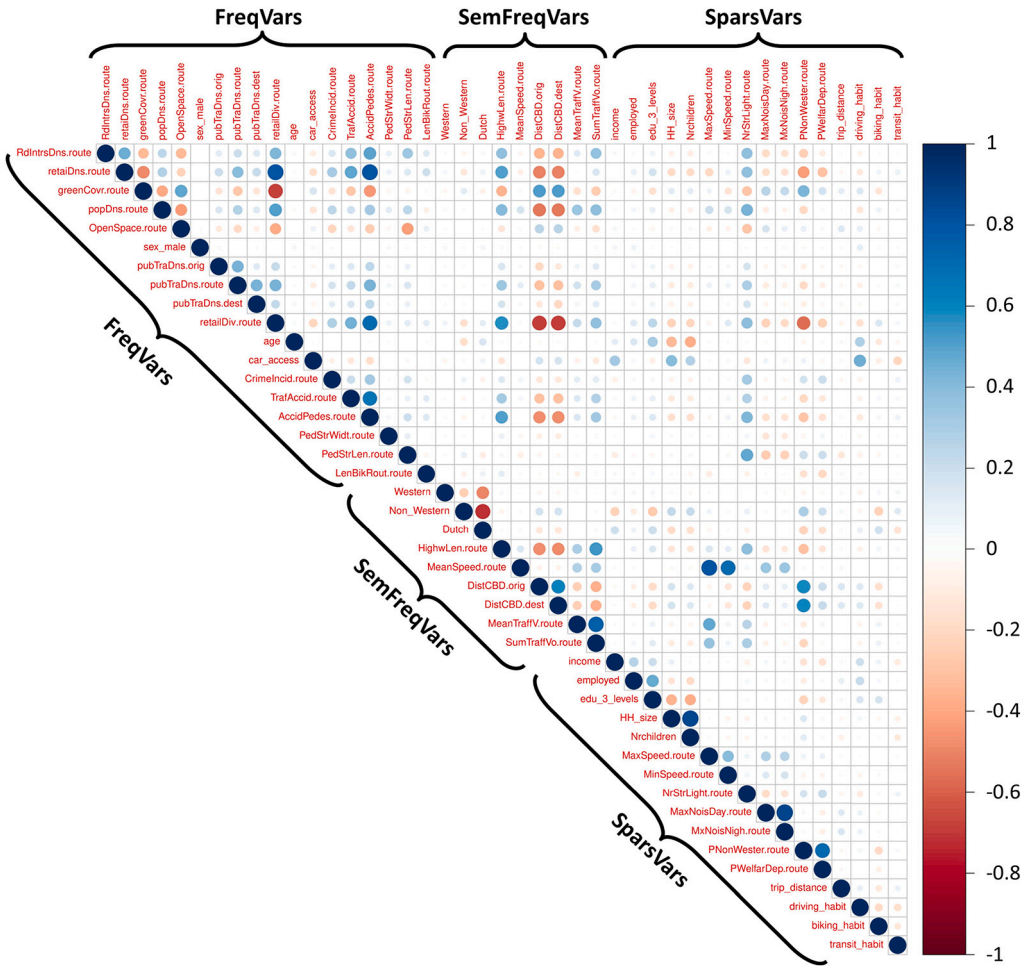


Fig. 7. Correlation Matrix of Behavior Determinants.

hinting at underlying spatial segregation, as well as retail diversity (coefficient of -0.56), which highly correlates with distance to CBD itself. Moreover, maximum and minimum speed limits along the route correlate with mean speed limits along the route (coefficients of 0.8 and 0.7 respectively). Other than that there are only correlations between sparsely studied variables. In other words, there is indeed added information in the sparsely studied variables.

We consequently trained decision trees to model mode choice behavior using the different variable subsets. Fig. 8 displays the performance results across 10 model runs for the three subsets for the decision tree (a) including the intermediate step of PCA reduction of variables with at least one correlation above 0.7 and (b) without the PCA step. One can see that the semi-frequently studied variables do not add very much to the predictive power of the decision tree, since the average weighted-average F1-score across multiple runs in the validation sample only increases by 0.01 in the PCA implementation and by 0.00 in the implementation without PCA reduction. In contrast to that, the sparsely studied variables improve the model significantly. The weighted-average F1 on the validation sample increases by 0.22 in the PCA methodology and by 0.19 in the methodology without PCA. Fig. 8 shows the F1 score for each mode of transport category when including all variables and using the PCA approach. One can see that biking is best predicted (see validation sample result), followed by walking, then public transit and finally driving.

Finally, Fig. 9 shows that the top four most important variables are sparsely studied variables, while many of the most studied variables are not actually that relevant.[13] The thickness of the edges is determined by the variable importance. One can see an upward trend of thick edges from the left bottom (sparsely studied variables) to the right top (most important variables), and a downward trend of thin edges from the left up (most studied variables) ending up in the right bottom part of the ranking. Strikingly, the most studied variable (Road Intersection Density/Street Connectivity) is only ranked 25th of 43 variables in terms of variable

[13] It is important to note that these results are based only on the extracted meta-analyses and systematic reviews and only on the evidence subset that relates to walking, biking, driving or public transit. Trip distance is a very common variable in transport mode choice models. Trip distance was significant in more than one study for behavior choice options that we did not model, e.g. “home-to-work distance” has been significantly related to active travel in general, and “distance to school” to “active school transport”.

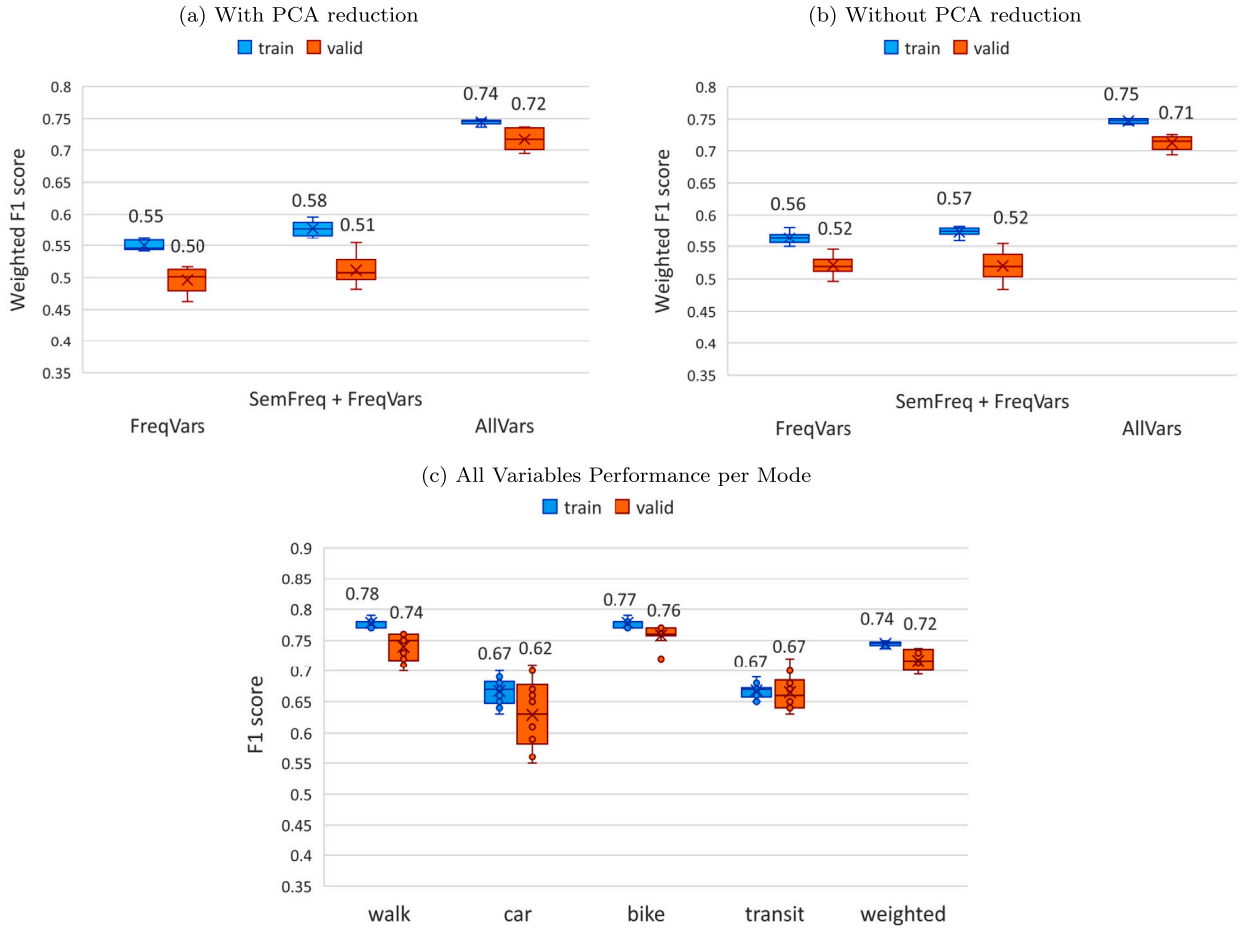


Fig. 8. Variable Subset Decision Tree Performance.

importance. This demonstrates that, as a matter of fact, most relevant variables are studied most seldom. The reason may lie in biased variable choices made by the authors of these studies.

6. Discussion and conclusion

We have presented a method for automated knowledge extraction of evidence on behavior determinants and interactions for different study groups and tested it using an example implementation for transport mode choice. The suggested method has several benefits for structural model validation. Our method is scaleable, saving the time needed to read and understand all the literature. There is also a gain in reproducibility, since knowledge extraction and synthesis happen in a standardized manner, reducing the risk of subjectivity bias. Moreover, the efficiency gain allows for new strategies and workflows in the scientific process. The automation of knowledge extraction and synthesis removes barriers to using existing empirical evidence and hence has the potential to improve the reuse of knowledge and improve the validity of new scientific work.

The potential of the knowledge graph extraction method for structural model validation was tested by applying the contained knowledge to a transport behavior prediction task. Our model implementation shows that sparsely studied significant variables can be of major importance to obtaining valid behavioral models. Less complete and conservative modeling approaches therefore are risking a severe variable selection bias. It is therefore crucial to have a comprehensive overview of all the literature. By extracting, synthesizing, and utilizing evidence on the determinants of a behavior, we can effectively reduce this bias: Moreover, we can better understand what the known unknowns and limitations of our models are if we do not have data to incorporate all significant variables.

However, this paper only presents a proof of concept of the suggested method. There is room for expanding on and improving the method. To channel future research efforts, below are a few suggestions:

- The training data for the BERT model can be improved by including more and more diverse articles. It is theoretically possible to train a generic BERT model that extracts independent and dependent variables, moderators, and study samples or objects for all kinds of disciplines. The availability of such a pretrained model would have the advantage that researchers don't have to

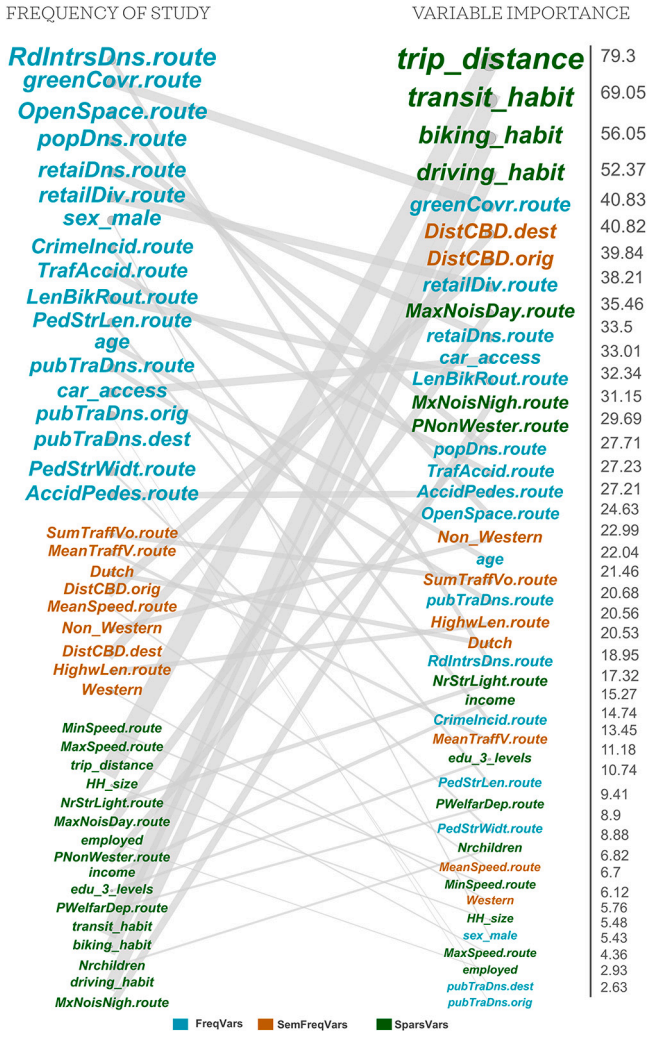


Fig. 9. Frequency of Study versus Variable Importance.

train their own algorithms to use this method. This requireslabeling a large set of articles from a diversity of disciplines with corresponding tags. To build a specific knowledge graph, a researcher would subsequently have to feed a subset of articles on a topic into the BERT model.

- The relation inference model proposed in this paper can be improved by incorporating other syntactical features. It was currently based on particular syntactical features, such as the Stanford dependency parser, which in itself involves occasional errors [92.2% accuracy [13]]. Despite the inclusion of 71 features, more work needs to be done to understand which other syntactical features could be relevant.
- As was pointed out in our modal choice method application, a major challenge is the semantic diversity of variable names, specifically identifying and processing synonymy, hyponymy, and polysemy. We implemented a preliminary solution by using semantic and string similarity measures as well as synonym/antonym analysis. As the performance metrics show, our approach works well, but more research may improve the variable harmonization method. This could be done e.g. based on transformer-based word embeddings (e.g. GPT-4 or BERT) to measure the semantic distance between entire phrases.
- Large Language Models (LLMs), such as GPT-4, Falcon, or LLAMA-2, may improve the NER task, the relation inference task and the variable harmonization task of our method framework. Future research should try and apply these newer models to these subtasks.
- Our proof of concept method only used meta-analyses and systematic reviews as evidence sources based on the rationale that they applied quality criteria themselves. However, a more inclusive approach would be to extent the BCDO with evidence attributes such as attributes of the methods used, the sample sizes, the reported uncertainty and the control variables, and then decide based on methodological criteria whether to include evidence instances in the knowledge graph. One can also add a more nuanced quality criterion to the BCDO that is computed based on these methodological attributes.

- The knowledge used to populate the BCDO allows us to make use of the *reasoning* abilities of the ontology [25]. For example, one could infer that if multiple study groups that are identified by different classes of the same variable (e.g. men, women) have different directions of statistical association or diverging significance levels of a specific behavior determinant, then the variable defining the social group is a moderator for the behavior determinant. More research is required to understand the value and potential of using reasoning to create new knowledge and implement causal theory to make causal inferences based on the collected knowledge [45].

Despite room for technical improvement, our method shows that it is feasible to automatically extract evidence instances from literature using NLP-based deep learning on a high level of quality, which can be used to validate the structure of simulation and projection models. We trust that our method could be used to validate the model structure of other phenomena or disciplines as well, provided that there is a large enough evidence base and manually extracting knowledge would be difficult to achieve and inefficient.

We hope that more research will follow and that the openly provided tools published alongside this paper can be used and improved to lead to a reliable, open-access method for automated knowledge extraction and model validation. With many societal challenges to be solved, the importance of reusing evidence for scenario modeling and explicit simulations remains critical. While the number of simulation methods is growing [e.g., [34], and so is their use [26], it is time for validation methods to catch up.

CRediT authorship contribution statement

Tabea S. Sonnenschein: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **G. Ardine de Wit:** Conceptualization, Writing – review & editing, Methodology, Supervision. **Nicolette R. den Braver:** Conceptualization, Writing – review & editing. **Roel C.H. Vermeulen:** Funding acquisition, Resources, Supervision, Writing – review & editing. **Simon Scheider:** Conceptualization, Formal analysis, Investigation, Methodology, Supervision, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

We have shared the populated knowledge graph and all code under this link, as also indicated in the manuscript: https://knowledgesynth.github.io/ontologies/bcdo.html.

Funding acknowledgment

This work was supported by EXPANSE and EXPOSOME-NL. EXPANSE has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 874627 and is coordinated by Utrecht University. EXPOSOME-NL is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.004.017).

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ins.2024.120232 (Ref. [44]).

References

[1] C. Abraham, S. Michie, A taxonomy of behavior change techniques used in interventions, Health Psychol. 27 (3) (2008) 379–387, https://doi.org/10.1037/0278-6133.27.3.379.

[2] I. Ajzen, The theory of planned behavior, in: Theories of Cognitive Self-Regulation, Organ. Behav. Hum. Decis. Process. 50 (2) (1991) 179–211, https://doi.org/10.1016/0749-5978(91)90020-T.

[3] F.N. Al-Aswadi, H.Y. Chan, K.H. Gan, Automatic ontology construction from text: a review from shallow to deep learning trend, Artif. Intell. Rev. 53 (6) (2020) 3901–3928, https://doi.org/10.1007/s10462-019-09782-9.

[4] J. Almagor, A. Martin, P. McCrorie, et al., How can an agent-based model explore the impact of interventions on children's physical activity in an urban environment?, Health Place 72 (June 2021) 102688, https://doi.org/10.1016/j.healthplace.2021.102688.

[5] N. Alshammari, S. Alanazi, The impact of using different annotation schemes on named entity recognition, Egypt. Inform. J. 22 (3) (2021) 295–302, https://doi.org/10.1016/j.eij.2020.10.004.

[6] S.D. Angus, B. Hassani-Mahmooei, "Anarchy" reigns: a quantitative analysis of agent-based modelling publication practices in JASSS, 2001-2012, J. Artif. Soc. Soc. Simul. 18 (4) (2015) 2001–2012, https://doi.org/10.18564/jasss.2952.

[7] L. Aston, G. Currie, A. Delbosc, et al., Exploring built environment impacts on transit use–an updated meta-analysis, Transp. Rev. 41 (1) (2021) 73–96, https://doi.org/10.1080/01441647.2020.1806941.

[8] C.A. Aumann, A methodology for developing simulation models of complex systems, Ecol. Model. 202 (3–4) (2007) 385–396, https://doi.org/10.1016/j.ecolmodel.2006.11.005.

[9] J. Badham, E. Chattoe-Brown, N. Gilbert, et al., Developing agent-based models of complex health behaviour, Health Place 54 (January 2018) 170–177, https://doi.org/10.1016/j.healthplace.2018.08.022.

[10] W. Borst, Construction of Engineering Ontologies, Ph.D. thesis, University of Twente, Enschede, the Netherlands, 1997.

[11] Centraal Bureau voor de Statistiek (CBS), Microdata onderzoek Onderweg in Nederland - ODiN 2019, 2019, https://www.cbs.nl/nl-nl/onze-diensten/maatwerk-en-microdata/microdata-zelf-onderzoek-doen/microdatabestanden/odin2019-onderweg-in-nederland-2019.

[12] E. Cerin, A. Nathan, J. van Cauwenberg, et al., The neighbourhood physical environment and active travel in older adults: a systematic review and meta-analysis, Int. J. Behav. Nutr. Phys. Act. 14 (1) (2017) 1–23, https://doi.org/10.1186/s12966-017-0471-5.

[13] D. Chen, C.D. Manning, A fast and accurate dependency parser using neural networks, in: EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference (i), 2014, pp. 740–750.

[14] A. Chiche, B. Yitagesu, Part of speech tagging: a systematic review of deep learning and machine learning approaches, J. Big Data 9 (1) (2022), https://doi.org/10.1186/s40537-022-00561-y.

[15] C. Choi, M. Cho, E.Y. Kang, et al., Travel ontology for recommendation system based on semantic Web, in: 8th International Conference Advanced Communication Technology, ICACT 2006 - Proceedings, vol. 1, 2006, pp. 624–627.

[16] A. Crooks, A. Heppenstall, N. Malleson, Agent-Based Modeling, vol. 3, Elsevier, 2017.

[17] J. Devlin, M.-W. Chang, K. Lee, et al., 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186.

[18] R. Ewing, R. Cervero, Travel and the built environment: a meta-analysis, J. Am. Plan. Assoc. 76 (3) (2010) 265–294, https://doi.org/10.1080/01944361003766766.

[19] C. Fellbaum (Ed.), WordNet: An Electronic Lexical Database, Language, Speech, and Communication, MIT Press, Cambridge, MA, 1998.

[20] N. Ferguson, Capturing human behaviour, Nature 446 (7137) (2007) 733, https://doi.org/10.1038/446733a.

[21] L. Festinger, A Theory of Cognitive Dissonance, Stanford University Press, 1957.

[22] A. Gangemi, V. Presutti, Ontology design patterns, in: Handbook on Ontologies, May, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 221–243.

[23] M. Georgeff, B. Pell, M. Pollack, et al., The belief-desire-intention model of agency, in: J.P. Müller, A.S. Rao, M.P. Singh (Eds.), Intelligent Agents V: Agents Theories, Architectures, and Languages, Springer Berlin Heidelberg, Berlin, Heidelberg, 1999, pp. 1–10.

[24] C. Goutte, E. Gaussier, A probabilistic interpretation of precision, recall and f-score, with implication for evaluation, in: D. Losada, J. FernandezLuna (Eds.), Advances in Information Retrieval, Microsoft Res; SHARP; ERCIM; CEPIS; BCS. 27th European Conference on Information Retrieval Research (ECIR 2005), Univ. Santiago Compostele, Tech. Sch. Engn., Santiago Compostela, Spain, March 21–23, 2005, in: Lecture Notes in Computer Science, vol. 3408, 2005, pp. 345–359.

[25] N. Guarino, D. Oberle, S. Staab, What is an ontology?, in: S. Staab, R. Studer (Eds.), Handbook on Ontologies, May 2009, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 1–17.

[26] M. Gunshin, K. Doi, N. Morimura, Use of high-fidelity simulation technology in disasters: an integrative literature review, Acute Med. Surg. 7 (1) (2020), https://doi.org/10.1002/ams2.596.

[27] J. Hastings, S. Michie, M. Johnston, Theory and ontology in behavioural science, Nat. Hum. Behav. 4 (3) (2020) 226, https://doi.org/10.1038/s41562-020-0826-9.

[28] T.A. Hilland, M. Bourke, G. Wiesner, et al., Correlates of walking among disadvantaged groups: a systematic review, Health Place 63 (March 2020) 102337, https://doi.org/10.1016/j.healthplace.2020.102337.

[29] G.J. Hollands, G. Bignardi, M. Johnston, et al., The TIPPME intervention typology for changing environments to change behaviour, Nat. Hum. Behav. 1 (8) (2017) 1–9, https://doi.org/10.1038/s41562-017-0140.

[30] P. Janich, Logisch-pragmatische Propädeutik: ein Grundkurs im philosophischen Reflektieren, Velbrück Wiss, 2001.

[31] P. Janich, Was ist denn nun Wahrheit-ganz praktisch gesehen?, in: Kultur und Methode, Philosophie in einen wissenschaftlich geprägten Welt, 2006, pp. 186–209.

[32] M. Katsumi, M. Grüninger, Choosing ontologies for reuse, Appl. Ontol. 12 (3–4) (2017) 195–221, https://doi.org/10.3233/AO-160171.

[33] D. Koller, N. Friedman, Probabilistic graphical models: principles and techniques, in: Adaptive Computation and Machine Learning, MIT Press, 2009.

[34] G. Lamé, R.K. Simmons, From behavioural simulation to computer models: how simulation can be used to improve healthcare management and policy, BMJ Simul. Technol. Enhanc. Learn. 6 (2) (2020) 95–102, https://doi.org/10.1136/bmjstel-2018-000377.

[35] K.R. Larsen, S. Michie, E.B. Hekler, et al., Behavior change interventions: the potential of ontologies for advancing science and practice, J. Behav. Med. 40 (1) (2017) 6–22, https://doi.org/10.1007/s10865-016-9768-0.

[36] Y. Li, T. Du, J. Peng, Understanding out-of-home food environment, family restaurant choices, and childhood obesity with an agent-based Huff model, Sustainability 10 (5) (2018), https://doi.org/10.3390/su10051575.

[37] E. Loper, S. Bird, Nltk: the natural language toolkit, CoRR, arXiv:cs.CL/0205028, 2002.

[38] F. Lovett, Rational choice theory and explanation, Ration. Soc. 18 (2006) 237–272, https://doi.org/10.1177/1043463106060155.

[39] S. Michie, M. Johnston, Optimising the value of the evidence generated in implementation science: the use of ontologies to address the challenges, Implement. Sci. 12 (1) (2017) 10–13, https://doi.org/10.1186/s13012-017-0660-2.

[40] S. Michie, M. Richardson, M. Johnston, et al., The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions, Ann. Behav. Med. 46 (1) (2013) 81–95, https://doi.org/10.1007/s12160-013-9486-6.

[41] S. Michie, J. Thomas, M. Johnston, et al., The human behaviour-change project: harnessing the power of artificial intelligence and machine learning for evidence synthesis and interpretation, Implement. Sci. 12 (1) (2017) 1–12, https://doi.org/10.1186/s13012-017-0641-5.

[42] E. Norris, A.N. Finnerty, J. Hastings, et al., Identifying and Evaluating Ontologies Related to Human Behaviour Change Interventions: a Scoping Review, 2018, pp. 1–42.

[43] E. Norris, A.N. Finnerty, J. Hastings, et al., A scoping review of ontologies related to human behaviour change, Nat. Hum. Behav. 3 (2) (2019) 164–172, https://doi.org/10.1038/s41562-018-0511-4.

[44] T. Novack, Z. Wang, A. Zipf, A system for generating customized pleasant pedestrian routes based on openstreetmap data, Sensors 18 (11) (2018), https://doi.org/10.3390/s18113794.

[45] J. Pearl, D. Mackenzie, The Book of Why, Basic Books, New York, 2018.

[46] J. Savage, B. Vila, Human ecology, crime, and crime control: linking individual behavior and aggregate crime, Soc. Biol. 50 (1–2) (2003) 77–101, https://doi.org/10.1080/19485565.2003.9989066.

[47] B. Schmidt, Modelling of human behaviour the PECS reference model, Artif. Intell. (c) (2002) 13–18.

[48] J. Schulze, B. Müller, J. Groeneveld, et al., Agent-based modelling of social-ecological systems: achievements, challenges, and a way forward, J. Artif. Soc. Soc. Simul. 20 (2) (2017), https://doi.org/10.18564/jasss.3423.

[49] S. Sharmin, M. Kamruzzaman, Association between the built environment and children's independent mobility: a meta-analytic review, J. Transp. Geogr. 61 (April 2017) 104–117, https://doi.org/10.1016/j.jtrangeo.2017.04.004.

[50] Z. Stavri, S. Michie, Classification systems in behavioural science: current systems and lessons from the natural, medical and social sciences, Health Psychol. Rev. 6 (1) (2012) 113–140, https://doi.org/10.1080/17437199.2011.641101.

[51] D. Tsatsou, E. Lalama, S.L. Wilson-Barnes, et al., NAct: the nutrition & activity ontology for healthy living, in: International Conference on Formal Ontology in Information Systems (FOIS), September 2021.

[52] U. Tuomainen, U. Candolin, Behavioural responses to human-induced environmental change, Biol. Rev. 86 (3) (2011) 640–657, https://doi.org/10.1111/j.1469-185X.2010.00164.x.
[53] T.J. Van Der Weele, P. Ding, Sensitivity analysis in observational research: introducing the E-Value, Ann. Intern. Med. 167 (4) (2017) 268–274, https://doi.org/10.7326/M16-2607.
[54] P. Vemer, I. Corro Ramos, G.A. van Voorn, et al., AdViSHE: a validation-assessment tool of health-economic models for decision makers and model users, PharmacoEconomics 34 (4) (2016) 349–361, https://doi.org/10.1007/s40273-015-0327-2.
[55] J. Verstegen, S. Scheider, Why the term prediction is overused, in: Spatial Data Science Symposium 2023 Short Paper Proceedings, Center for Spatial Studies, UC Santa Barbara, 2023.
[56] A. Voinov, F. Bousquet, Modelling with stakeholders, Environ. Model. Softw. 25 (11) (2010) 1268–1281, https://doi.org/10.1016/j.envsoft.2010.03.007.
[57] I. Vojnovic, A. Ligmann-Zielinska, T.F. LeDoux, The dynamics of food shopping behavior: exploring travel patterns in low-income Detroit neighborhoods experiencing extreme disinvestment using agent-based modeling, PLoS ONE 15 (2) (December 2021) 1–25, https://doi.org/10.1371/journal.pone.0243501.
[58] L. Whitmarsh, G. Seyfang, S. O'Neill, Public engagement with carbon and climate change: to what extent is the public 'carbon capable'?, Glob. Environ. Change 21 (1) (2011) 56–65, https://doi.org/10.1016/j.gloenvcha.2010.07.011.
[59] W. Winkler, String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage, Proc. Sect. Surv. Res. Methods (1990).
[60] A.J. Wright, S. Michie, E. Norris, et al., Ontologies relevant to behaviour change interventions: a method for their development, Wellcome Open Res. 5 (2020), https://doi.org/10.12688/wellcomeopenres.15908.3.
[61] Y. Yang, A.H. Auchincloss, D.A. Rodriguez, et al., Modeling spatial segregation and travel cost influences on utilitarian walking: towards policy intervention, Comput. Environ. Urban Syst. 51 (January 2015) 59–69, https://doi.org/10.1016/j.compenvurbsys.2015.01.007.