Methods

Data source and pre-processing

To develop the disease-specific language model, we accessed data on 1.9 million patients from the HealthVerity longitudinal repository. This included medical insurance claims data from an undisclosed provider, represented as ICD-10 codes, and physician notes in free-text form from Amazing Charts LLC. In compliance with HIPAA regulations, all data used in this study was de-identified, containing no personally identifiable information. As such, institutional review board (IRB) approval was not required for the initial setup of the database and the resulting study data, as the analysis of de-identified data does not constitute human subjects' research. Likewise, informed consent was not necessary due to the de-identified nature of the dataset.

Of the 1.9 million patients identified, data from 82,722 patients diagnosed with insomnia (Supplementary Table 1) and at least one physician note (486,885 visit notes in total) were considered and split into a model training set of 82,672 patients and a test set of 50 patients. Both sets were stratified by age and sex. Only information on the chief complaint, review of the system, physical exam, the assessment, and history of present illness were considered for model training. Data related to past medical history was discarded, as it may have contained information on diseases not related to insomnia.

Using a Simple Ontology as Model Input

The Insomnia Daytime Symptoms and Impacts Questionnaire (IDSIQ)8 was used as a starting point to define the disease-specific language model input (referred to as the simple ontology). IDSIQ is a patient-reported outcome instrument that consists of 14 items, an item being a sentence describing a daytime impairment symptom as perceived by patients, which were grouped into three domains (cognition, emotional, physical) by the insomnia experts (Supplementary Methods Insomnia Experts). While a different instrument or knowledge base could have been considered, the IDSIQ was chosen by insomnia experts because of its totality in describing symptoms of daytime impairment. To complement the patient-reported symptoms, two insomnia experts provided terms of daytime impairment commonly used in their clinical practice to describe the symptoms reported by the patients. While words such as sleepy, tired, drowsy, and fatigue9 have different meanings to patients, the same cannot be assumed of clinicians who are not experts in insomnia. To account for the potential heterogeneous language used by non-sleep specialists, the terms provided by the insomnia experts - along with the original IDSIQ items - were used as input (referred to as the simple ontology) to find contextually similar words through the DiSMOL framework (Fig. 1, Supplementary Data 1).

Learning a disease specific model

DiSMOL used Word2vec10 and hyperparameter optimization. The hyperparameter space (Supplementary Table 3) was restricted to a recommended range11, and the models were optimized using a metric inspired by Beam and colleagues12. The metric involved creating a list of insomnia symptom pairs with a known relationship (Supplementary Data 4), assigning a semantic type to each token, randomly drawing 10,000 tokens of the semantic supertype, daytime impairment (Supplementary Table 5), and calculating the cosine similarity for each pair of tokens. The resulting distribution of cosine similarities was used to evaluate the models, with the metric being the fraction of known relationships within the 95th percentile cosine distance distribution.

Eight models, trained on different bootstrap samples each using the best performing hyperparameter configuration, were used to create an ensemble model. This involved averaging distances between tokens across the eight models to reduce variations in close neighbors due to small changes in the data and initialize the weights of the model. To assess the suitability of pre-trained language models versus disease-specific models, popular models such as Bidirectional Encoder Representations from Transformers (BERT) were evaluated alongside the disease-specific Word2Vec ensemble model. The disease-specific model (Supplementary Table 2) outperformed BERT, in median performance across 5 independent runs by 83% (Table 1), enhancing the performance for downstream tasks.

Ontology enrichment through disease specific model

The simple ontology was subsequently enriched by retrieving additional terms from the disease-specific language model. Only pairs (x: an item or corresponding insomnia expert synonym, y: retrieved word2vec term) within the 0.963 percentile with respect to the distribution of cosine distances were retrieved, and only those with the semantic supertype of daytime impairment were kept (Supplementary Table 5). Furthermore, the words needed to appear in at least 0.24% of all patients and have a clear semantic directionality as determined by the medspaCy ConText algorithm[13]. The remaining set of 425 terms was further investigated by insomnia experts and words not describing daytime impairment were removed resulting in 109 words (Supplementary Data 2). We acknowledge that the validation rate of these terms was approximately 25%. Regarding the terms that were ultimately discarded, our analysis shows that while they did not meet the criteria for inclusion in the final dataset, many were loosely related to the domain of interest but lacked the necessary specificity or direct relevance to daytime impairment as defined for our study. To clarify, the system was not designed for fully automated use; instead, it was developed to aid experts by providing a broader array of potential terms for further scrutiny. Finally, the daytime impairment ontology was examined for lexical consistency (i.e., words with the same lemma were present in all or no domains).

By following the DiSMOL approach, the language of insomnia experts was enriched by incorporating terminology used by a large sample of healthcare providers and non-sleep experts, without needing to conduct an extensive and time-consuming questionnaire. Furthermore, as this approach uses physician notes written by healthcare providers prior to using these notes for research purpose, there is no potential bias with respect to the

Fig. 1 | Disease-Specific Medical Ontology Learning (DiSMOL) Framework. The disease-specific language model is trained on physician notes. Known symptoms, as informed by insomnia experts or existing instruments form the simple ontology, which is used as an input to the model to find contextually similar words. The resulting enriched ontology is then validated by an expert panel. Subsequently, the validated enriched ontology is utilized to transform physician notes into structured data, enabling the investigation of the disease burden.

Hawthorne Effect and therefore provides a representation of language use in everyday clinical practice. Lastly, owing to minimal human intervention (Fig. 1), the approach highlights scalability and applicability to other disorders.

Inference of the daytime impairment ontology on text data

The ontology was utilized to identify terms from physician notes, representing the presence of a daytime impairment for a patient at a specific encounter. The medspaCy ConText algorithm was utilized to measure the correct semantic directionality of terms. The presence of a daytime impairment symptom was only confirmed if a word had the correct semantic directionality. For instance,“being happy” would not constitute an impairment, but“not being happy” would be identified as one. The presence of an impairment was assessed on the level of a patient encounter with a healthcare provider. An encounter was considered to have an impairment if one or more daytime impairment symptoms were present. The number of impairments for a specific encounter was not considered.

Ontology evaluation

The accuracy of the enriched insomnia daytime impairment ontology was evaluated against two baseline representations based on ICD-10 codes. The first representation (ICD-clin) used a set of ICD-10 codes selected by insomnia experts (Supplementary Table 4). For the second representation (ICD-clin-DiSMOL), insomnia experts matched ICD-10 codes with symptoms generated from DiSMOL framework (Supplementary Data 3).

To generate our ground truth data, an insomnia expert annotated 108 physician notes from the test set of 50 patients according to the presence or absence of physician-reported daytime impairment symptoms in the three domains: cognition, emotional, and physical. The ability of ICD-clin, ICD-clin-DiSMOL, and DiSMOL to match ground truth data was evaluated using the sensitivity, precision, and F1-score14

.

Treatment effect estimation

To compare the ability of ICD-clin, ICD-clin-DiSMOL, and DiSMOL to quantify the effect of treatment on the number of daytime impairment symptoms in individuals with insomnia, three distinct patient populations were identified based on the treatment administered: trazodone, benzodiazepine consisting of estazolam, flurazepam, lorazepam, quazepam, temazepam, triazolam), and non-benzodiazepine receptor agonists (non-BzRAs) consisting of eszopiclone, zaleplon, zolpidem. Of the 82,722 patients originally identified, only those with an insomnia diagnosis within 6 months prior to treatment start and aged ≥18 years at baseline were included. Exclusion criteria included an insomnia diagnosis date >6 months before treatment start, palliative care, active malignancies, or pregnancy. Patients were followed retrospectively for 182 days (6 months) prior to their first pharmacological treatment for insomnia and for an average of three months during treatment. To estimate treatment effect, the within-subject relative difference in the number of daytime impairment symptoms prior to and during treatment with trazodone (n = 1522 patients), benzodiazepine (n = 1045 patients), and non-BzRAs (n = 2361 patients) were measured. A censoring of the treatment start day was applied for the calculation of the treatment effects. The statistical significance of the difference in outcome between the pre-treatment period and post-treatment period was calculated using a two-sided t-test. Multiple hypothesis tests using the false discovery rate (FDR) corrections with the Benjamini–Hochberg procedure15 were applied. The effect estimation was calculated by counting the number of reported symptoms in 100 patient-years and subtracting the number of symptoms in the baseline period reported in 100 patient-years. CIs <0.05 were considered statistically significant.

Although the baseline characteristics of the three study populations were comparable, except for psychiatric co-morbidities, which were less frequent in the non-BzRA group (Table 2), the study design did not allow for a comparison between the different treatments, as the control for confounders only applies within a specific population.

Statistics and reproducibility

All statistical tests used, the sample sizes, the number of replicates, and how replicates were defined are described in the methods section under data source and preprocessing, learning a disease specific model, ontology enrichment through disease specific model, ontology evaluation, and treatment effect estimation.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.