Perspective: Leveraging Domain Knowledge for Tabular Machine Learning in the Medical Domain

Arijana Bohr[1], Thomas Altstidl[1], Bjoern Eskofier[1,2] and Emmanuelle Salin[1]

[1]Machine Learning and Data Analytics Lab,
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Germany
[2]Institute of AI for Health, Helmholtz Zentrum München
German Research Center for Environmental Health, Neuherberg, Germany
{firstname.lastname}@fau.de

Abstract

There has been limited exploration of how domain knowledge can be effectively integrated into machine learning for medical tabular data. Traditional approaches often rely on non-generalizable processes tailored to specific datasets. In contrast, recent advances in deep learning for language and tabular data are leading the way toward more generalizable and scalable methods of domain knowledge inclusion. In this paper, we first explore the need for domain knowledge in medical tabular data, categorize types of medical domain knowledge, and discuss how each can be leveraged in tabular machine learning. We then outline strategies for integrating this knowledge at various stages of the machine learning pipeline. Finally, building on recent advances in tabular deep learning, we propose future research directions to support the integration of domain knowledge.

1 Introduction

Tabular data plays a fundamental role in the medical field, capturing patient-specific details such as demographics, medical history, biomarkers, and diagnostic codes (Mao et al., 2024). Many clinical machine learning models rely on this data for tasks such as disease diagnosis (Ahsan et al., 2022) and adverse events prediction (Tomašev et al., 2021).

However, developing these models poses unique challenges. For instance, models can often learn shortcuts when modeling the data, leading to potentially harmful decisions. Caruana et al. (2015), for example, show that a model trained to predict pneumonia risk can incorrectly identify asthma as a protective factor. This error can occur because asthmatic patients generally receive more aggressive treatment, leading to better outcomes.

In contrast to clinicians who draw on prior training and domain expertise, models are typically developed with limited prior knowledge (Moor et al., 2023). They rely on statistical associations between input features and targets and do not understand the underlying physiology (Moor et al., 2023). Learning these associations can be further complicated by the heterogeneous features and complex interactions present in medical datasets (Ruan et al., 2024).

The lack of knowledge can also hinder the development of models for specialized medical tasks (Moor et al., 2023), as it can limit their ability to perform reliably in various clinical settings. In addition, inconsistencies in data standardization of medical datasets (Ahmadian et al., 2011) can be a barrier to the generalizability of models across medical environments.

This paper explores how the integration of domain knowledge into machine learning for medical tabular data can help address these challenges. In particular, it can guide variable selection (Wu et al., 2022), mitigate data quality issues (Curé, 2012) and help establish consistent standardization (Shi et al., 2021). It can also help ensure that models meet natural laws and regulatory requirements, which data-driven approaches may ignore (Von Rueden et al., 2021). Ultimately, this could support the translation of machine learning into clinical practice, a hurdle many existing models have yet to overcome (El Naqa et al., 2023).

Despite the widespread use of tabular data in healthcare, to our knowledge, there has been no comprehensive investigation of domain knowledge integration for medical tabular data. In this paper, we first detail the types of medical domain knowledge and their potential uses. We then provide an overview of strategies for incorporating medical domain knowledge into tabular machine learning at all pipeline stages. In particular, we investigate how recent methods in table representation learning, such as foundation models (Hollmann et al., 2023a) or LLM-based table representation (Sui et al., 2024), can be adapted for this purpose. Finally, we suggest promising research directions

Proceedings of the 4th Table Representation Learning Workshop, pages 143–155
July 31, 2025 ©2025 Association for Computational Linguistics

for automated knowledge integration in clinical machine learning for medical tabular data.

2 Related Works

Domain knowledge encompasses relevant information about the machine learning task, including relevant features, taxonomies, logical constraints, and probability distributions (Dash et al., 2022). It is also referred to as background or prior knowledge. Domain knowledge has been incorporated into various fields of machine learning, such as physics and engineering, where it is used to combine data with mathematical and physics-based models (Karniadakis et al., 2021; Willard et al., 2022).

In the medical domain, the importance of integrating domain knowledge has been increasingly recognized (Mao et al., 2024; Leiser et al., 2023; Von Rueden et al., 2021), especially in areas such as medical imaging (Xie et al., 2021). While previous work has shown that domain knowledge can benefit tabular clinical decision systems (Sirocchi et al., 2024), it is often poorly integrated into clinical machine learning pipelines and requires custom algorithms (Sirocchi et al., 2024).

Xie et al. (2021) identify three challenges hindering the adoption of domain knowledge in medical computer vision models, which are also relevant to tabular data: identifying relevant sources, selecting appropriate representations, and integrating them into deep learning models.

3 Medical Domain Knowledge

In this section, we build on prior work in machine learning and domain-informed models (Von Rueden et al., 2021; Mao et al., 2024) to propose a categorization of medical domain knowledge.

3.1 Patient Data

Definition Patient data encompasses a wide range of health-related information, such as demographics, laboratory values, and vital signs. These data are commonly stored in systems like Electronic Health Records (EHRs).

The accessibility of patient datasets can vary considerably. MIMIC (Johnson et al., 2023) or UK Biobank (Sudlow et al., 2015) are available to researchers through application procedures, while most datasets are only accessible within individual institutions. These datasets may reflect the biases of specific patient populations. Other sources, such as population-wide health statistics, from initiatives like the Global Burden of Disease (Vollset et al., 2024), can provide context to assess generalizability. In addition, knowledge graphs can be developed from datasets such as cancer registries to understand the variation in outcomes (Hasan et al., 2019). Furthermore, biomedical databases that capture gene-gene or protein-protein interactions encode biological relationships and can serve as prior knowledge to inform downstream model training and inference (Wysocka et al., 2023).

Representation Patient data is often represented by datasets of various modalities that can be used to train or pre-train medical models.

Integration Patient data can be used for training and subgroup analyses, bias detection, and generalizability evaluation across diverse cohorts. Patient statistics can also inform feature engineering.

3.2 Formal Knowledge

Definition Formal knowledge encompasses established biomedical and scientific information recognized by scientific consensus. It originates from authoritative sources, such as medical textbooks or clinical guidelines, which can establish standardized procedures for clinical practice.

Formal knowledge can be *quantitative*, often represented through mathematical models that estimate biomarker dynamics or disease progression, such as pharmacokinetic models of drug absorption (Lin and Wong, 2017) or tumor growth models (Albano and Giorno, 2006; Tabatabai et al., 2005). Known clinical thresholds (e.g., defining sinus tachycardia as heart rate $\geq$ 100 bpm at rest (Page et al., 2016)) can guide data encoding and interpretation. Additionally, quantitative rules support data quality control by flagging physiologically implausible values.

Formal knowledge can also be *qualitative*, capturing the known interactions of patient characteristics. For instance, diagnosing delirium relies on behavioral and cognitive changes assessed through mental status exams (Tieges et al., 2018). Similarly, clinical gestalt refers to the ability of a physician to synthesize signals such as facial expressions or posture to form early diagnostic impressions (Cramer et al., 2025). Though laboratory tests often confirm a diagnosis, initial suspicion can stem from these assessments, such as hyperpigmentation in vitamin B12 deficiency (Brescoll and Daveluy, 2015).

Representation Formal knowledge can be represented as rules, lookup tables (e.g., scoring ranges, reference intervals), and flow charts or other categorical mappings for qualitative associations.

Integration Formal knowledge can be used for feature engineering, data cleaning, encoding medical relationships, integrating medical constraints, and validation.

3.3 Medical Semantics

Definition Medical semantics refers to standardized representations of biomedical concepts that support interoperability between datasets.

In tabular medical datasets, biomedical concepts are often expressed in varying forms, through free text and different coding systems. This variability can hinder the generalizability of machine learning models. To address this, semantic frameworks like SNOMED CT (Chang and Mostafa, 2021) and the Unified Medical Language System (UMLS) Lindberg et al. (1993) offer structured vocabularies and ontologies (Gaudet-Blavignac et al., 2021). LLMs can also generate medical semantic embeddings that enrich tabular data with contextual meaning. For example, Michalopoulos et al. (2021) introduce UmlsBERT, which incorporates domain knowledge from UMLS by linking terms with shared concepts and semantic types.

Representation Medical semantics can be represented through ontologies and dictionaries or captured by using biomedical language models.

Integration Medical semantics can be used for preprocessing, standardization, or to enrich existing data with semantic hierarchy or similarity.

3.4 Experimental Medical Findings

Definition Experimental medical findings derived from data analyses, clinical studies, or trials often reveal potential interactions between biomedical concepts, even if causal relationships are not yet established or still require scientific consensus. For example, current evidence from controlled exposure studies in children supports an association between adverse behavioral outcomes and synthetic food dye (Miller et al., 2022). Experimental findings are typically also compiled in clinical guidelines used by physicians. They are classified into multiple categories of recommendations (Class I, IIa, IIb, II and III) and levels of evidence (A, B, or C) (McDonagh et al., 2023). These findings can serve as hypotheses to guide the design of machine learning models.

While clinical guidelines can be difficult to interpret due to their length and variations in format (e.g., text, flowcharts, tables), advances in retrieval augmented generation models lead the way towards a more efficient extraction of relevant information (Kresevic et al., 2024).

Representation: Experimental findings can be represented as soft rules with confidence scores, probabilistic associations, or model priors.

Integration Experimental findings can be used to incorporate promising hypotheses that are supported by preliminary evidence. It may be used to explore feature relationships during feature engineering, prioritize variables during feature selection, and introduce soft constraints during model training or validation.

3.5 Professional Insights

Definition Reasoning developed by experienced clinicians provides essential context when interpreting information. With years of clinical experience, even limited data can be synthesized to make a diagnosis (Groves et al., 2003). This is demonstrated, for example, by optometrists outperforming novices in diagnosing glaucoma when data is limited (Ghaffar et al., 2025).

Expert insight is particularly valuable for identifying potential confounding factors when developing machine learning models for clinical use. For instance, patients nearing the end of life may establish legal directives, such as Do Not Resuscitate (DNR) orders, to limit medical intervention by their wishes (Schmidt et al., 2015). However, such directives are often not recorded in structured datasets and may be communicated only verbally.

Representation Professional insight can be formalized through rules, thresholds, or guidelines derived from expert interviews or consensus (e.g., expert surveys).

Integration Expert input can inform data collection through the design of study protocols and guide the selection and construction of features. It also plays a key role in validating models, interpreting outliers, and enabling feedback loops for iterative refinement.

Integrating Domain Knowledge for Multi-Label Post-operative Complication Prediction

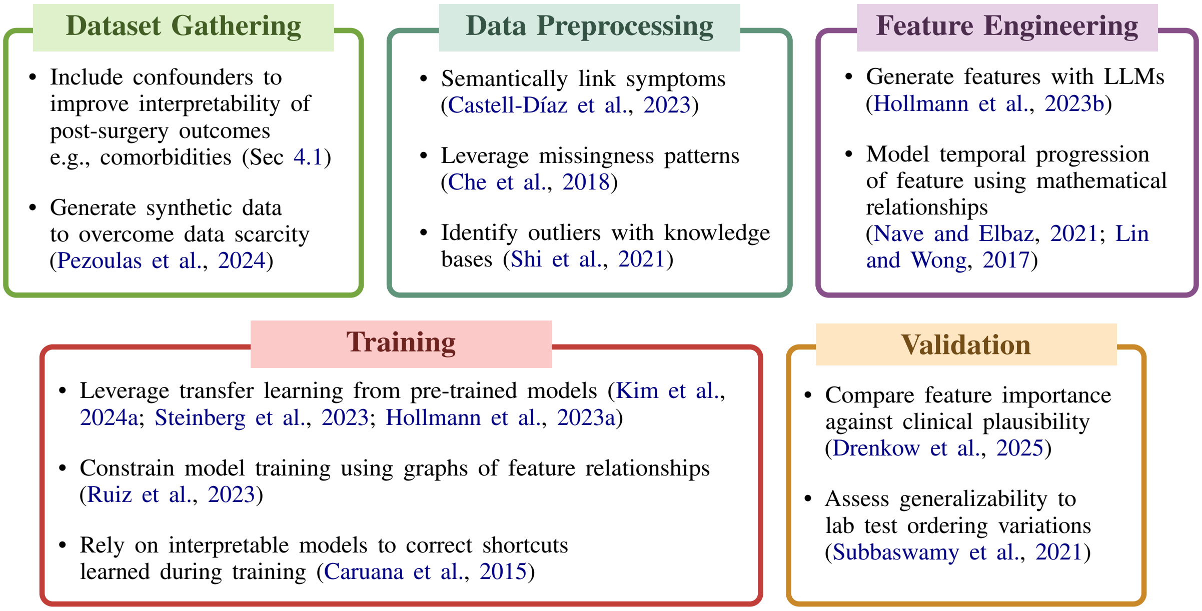

Figure 1: Possible integrations of domain knowledge for the use-case post-surgery complications prediction

4 Integrating Domain Knowledge

In Section 3, we explored the various forms of medical domain knowledge. Here, we examine each stage of the machine learning pipeline, from data collection to model validation, and highlight opportunities to meaningfully integrate domain expertise. We also focus on how advances in deep learning can be incorporated for domain knowledge integration and suggest promising research directions. In Figure 1, we provide an example of how domain knowledge can be integrated into the use case of post-surgery complications prediction.

4.1 Dataset Creation and Selection

Data collection Medical domain knowledge and professional insight are critical to data collection, especially in the case of *prospective studies*. Expert input (see Section 3.5) is essential when designing the study protocol, selecting data sources, defining patient populations, and determining which features to collect. Potential confounders should be considered during study design and data collection or assessed during analysis (Jager et al., 2008; Kahlert et al., 2017). A common strategy involves defining an a priori set of covariates to account for (Brookhart et al., 2010). For example, in a study investigating diabetes and ischemic heart disease, researchers could control for age by including only participants over 65 (Jager et al., 2008).

Beyond addressing confounders, incorporating additional relevant variables can help capture clinical context. Savchenko et al. (2023), for example, incorporate patient socio-demographic information to model the clinical dynamics of non-invasive bladder cancer treatment. Their inclusion yields an 8.14% performance gain over the baseline model lacking these features (Savchenko et al., 2023).

For *retrospective studies*, leveraging public datasets can also enrich training data. Factors such as demographic statistics can help select appropriate datasets. Ontologies can also be used to semantically categorize features, enabling table comparisons (Woźnica et al., 2024).

Synthetic data Synthetic data can help protect patient privacy or increase data size (Pezoulas et al., 2024). Bayesian networks can be used to generate synthetic patient data by modeling probabilistic relationships and latent variables (Tucker et al., 2020). These relationships can be informed by expert knowledge (Rabaey et al., 2024) or learned from existing datasets (Tucker et al., 2020). To ensure that the generated data maintains strong inferential properties, informative prior knowledge is essential to appropriately weight the different network structures (Young et al., 2009). Simulation-based methods can also leverage domain knowledge to generate data points. Deist et al. (2019) propose a technique that integrates prior knowledge using domain-informed kernels. The method performs well in low-data, high-dimensional settings but is surpassed by data-driven approaches as training data increases. Shi et al. (2022), for instance, show that when data-driven methods use large amounts of data, they can generate synthetic data that closely resembles real data.

Large language models have also been proposed for synthetic data generation (Zhang et al., 2023). However, this approach should be further tested in the medical domain in terms of privacy preservation. Kim et al. (2024b) propose combining LLMs with attribute constraints to generate synthetic financial data. Yet, they notice that using constraints could reduce diversity in some attributes, which may cause issues for data with high variability. These findings may also be relevant for similar approaches in the medical domain.

While synthetic data is often used to replace or complement training data, it can also help train tabular models. TabPFN (Hollmann et al., 2023a), a transformer-based model for tabular tasks, is trained on a large number of synthetic datasets, reducing reliance on sensitive real-world data. Recent work has demonstrated that domain knowledge can improve its adaptability to specific data types. For example, Perciballi et al. (2024) enhanced TabPFN's performance on metagenomic data by modifying the generative model priors to better reflect the sparsity and variability of this domain. However, the high variability in their results indicates that further experimentation is needed.

Future Research When working with a small dataset, a common strategy is to identify semantically or structurally similar datasets that can be leveraged through transfer learning. Advances in semantic data type detection (e.g., Hulsebos et al. (2023)) could lead to more informed dataset selections when combined with medical ontologies.

Synthetic data offers another promising research avenue for bias mitigation and data augmentation. The explicit inclusion of domain knowledge could guide this process, especially for low-resource domains. However, more research is still needed to compare the various methods of synthetic data generation in terms of privacy preservation, fidelity, bias, and clinical relevance.

4.2 Data Preprocessing

Cleaning Clinical data often contains inconsistencies that require tailored preprocessing. While such issues are best mitigated through standardized data collection protocols, missing data and non-standardized entries remain common and are sometimes unavoidable.

Numerical values suffer from inconsistent units due to varying practices across laboratories and general practitioners (e.g., 'g/dL', '??', 'NULL') (Shi et al., 2021). Domain knowledge can guide semantic alignment and harmonization through the identification of valid unit conversions or the correction of implausible entries (e.g., checking whether values are in acceptable ranges). For instance, Shi et al. (2021) automatically derive conversion rates, detect outliers, and identify extreme ranges using literature and knowledge bases.

Categorical values also require standardization. For this, medical knowledge bases can provide structured vocabularies (Chang and Mostafa, 2021; Bodenreider, 2004), and dictionaries can define permissible value labels, helping flag and correct invalid entries (Pilowsky et al., 2024). Beyond rule-based methods, ontology embedding techniques can leverage clinical ontologies to generate vector representations of terms (Zahra and Kate, 2024; Castell-Díaz et al., 2023). These embeddings enable the suggestion of the semantically related post-coordinated expression (Castell-Díaz et al., 2023).

Using LLMs for automated tabular data cleaning could alleviate the need for tailor-made outlier detection and error correction algorithms (Bendinelli et al., 2025). However, (Bendinelli et al., 2025) observe that LLMs tend to use brute force for data cleaning. Providing contextual knowledge, such as partial guidance on how to correct an error, often improves the results.

Missing data A common approach to handling missing data is complete case analysis, which excludes patients with incomplete information. This can introduce selection bias when missingness is related to underlying clinical factors (Haneuse, 2016). Clinical insight is therefore essential to assess if missingness is occurring at random. In the case of longitudinal data, missingness patterns can be especially informative (Che et al., 2018). For instance, stable patients may have specific lab tests omitted (Raebel et al., 2016), or patients experiencing severe toxicity may be more likely to drop out of a clinical trial (Bell et al., 2014).

Medical context also informs the design of imputation strategies. Multi-omics correlations from external datasets can, for instance, help impute genetic data (Lin et al., 2016). More recently, LLM-based imputation methods have shown significant improvement over baselines for data ‘missing not at random’ (Hayat and Hasan, 2024).

Future Research Preprocessing is crucial for ensuring interoperability, especially when combining datasets from multiple institutions where data quality often varies. In particular, poor standardization across datasets and a high rate of missing data impact the quality of tabular medical datasets. Current initiatives on the interoperability of healthcare databases aim to lessen the need for custom preprocessing (Semler et al., 2018).

Recent advances in table understanding methods that identify the semantic and syntactic types of cells (Zhang et al., 2020; Sun et al., 2021) represent a promising step toward developing end-to-end pipelines for automatic clinical data preprocessing. Further research on the use of medical vocabularies or ontologies in conjunction with LLMs could improve semantic interoperability. More broadly, LLMs are a promising research direction for automatized data cleaning and standardization. However, to our knowledge, they have not yet been applied to medical datasets with complex feature interactions. Thus, further adaptation and validation of this method to such datasets is necessary.

Although numerous statistical imputation techniques exist, many rely on the assumption that data is missing at random. This assumption often fails to account for the clinical context behind missingness. There is a growing need for frameworks that can represent the reasons behind missing data to address data ‘missing not at random’. In cases where the underlying mechanisms can be known or approximated, mathematical models (e.g., pharmacokinetic models) could be leveraged to infer and impute specific features (Lin and Wong, 2017).

4.3 Feature Engineering

Feature selection and creation Domain knowledge is frequently integrated into feature selection, particularly in biomedical applications, where datasets often contain relatively few instances but many features. In this context, it can help reduce complexity and enhance model performance. The effectiveness of this approach depends on the use of accurate and contextually appropriate knowledge: Wu et al. (2022) show that well-curated, targeted domain knowledge yields superior results compared to indiscriminate application.

Domain knowledge can also be used to generate new features from existing ones. Features can be handcrafted based on clinical knowledge and, in particular, mathematical relationships. Nave and Elbaz (2021) train a machine learning model to predict tumor size over time. Their results showed that adding mathematical model outputs significantly improved performance: their tumor size prediction accuracy increased from 72.5% to 86.33%.

Hollmann et al. (2023b), on the other hand, use LLMs to engineer additional features automatically based on a dataset description. This approach can be further extended by integrating domain expertise. For example, an estimation of medication absorption could be calculated using baseline patient information (Rajagopalan and Gastonguay, 2003).

Table serialization Clinical data can also be serialized into text and processed using language models. This can allow models to extract semantically rich representations that might not be apparent through standard tabular processing alone. For example, Chen et al. (2023) apply this approach to prognosis prediction, leveraging medical knowledge from pre-training data to enrich tabular representations. Similarly, Slack and Singh (2023) propose a pipeline that integrates domain knowledge into LLM-based differential diagnosis prediction. They enrich tabular data with disease-specific instructions and show that including this can often significantly increase performance.

Future Research Language models offer a promising avenue for the automated engineering of additional features based on domain knowledge. However, their outputs may introduce biases, as careful assessment of these methods is still needed. For instance, Küken et al. (2025) observe that LLMs often rely too heavily on simplistic operations, such as addition, when generating features. Including information on formal relationships from domain knowledge to engineer features could be a way to avoid this bias.

While LLMs have been used for medical tabular tasks, they have yet to be extensively tested on clinical datasets with high-dimensional features. Multimodal approaches combining a language model and high-dimensional table representation may be more appropriate (AlSaad et al., 2024). However, current research on such multimodal models is still limited. In addition, using LLMs for feature engineering also requires more extensive testing of the potential propagation of training data biases.

4.4 Training

Leveraging graph representations Domain knowledge can be used to introduce clinically meaningful inductive biases during training, guiding models to learn patterns that align with established medical understanding. Graph representations of domain knowledge can encode structured relationships. For instance, Middleton et al. (2024) jointly process tabular data and knowledge graphs to identify therapeutic genetic targets. Similarly, Ruiz et al. (2023) encode prior knowledge in a graph structure, influencing how feature connections are learned—demonstrating efficiency in high-dimensional, low-sample settings such as genomics. The hierarchical structure of medical concepts has also been incorporated into knowledge graphs to improve single-cell classification (Mojarrad et al., 2024).

Other architectures In physics-informed neural networks, regularization losses can enforce expected behavior in a model's outputs (Cuomo et al., 2022). For example, Nguyen et al. (2020) introduce a domain-specific loss function based on the dose volume histogram from radiation therapy. They show that this loss improves results across most evaluation categories (Nguyen et al., 2020).

Using interpretable models can also help interpret patterns and use domain knowledge to correct potential unwanted shortcuts that conflict with clinical reality. For instance, Caruana et al. (2015) develop generalized additive models with pairwise interactions for a pneumonia detection task. When the model incorrectly learns, for example, that asthma lowers the risk of pneumonia, it can be addressed by reshaping the learned effect function to reflect the correct association.

Foundation model pre-training Through self-supervised pre-training, models can leverage the longitudinal nature of EHRs. For example, Steinberg et al. (2023) pre-train a time-to-event transformer-based model from EHRs medical codes. This helps model medical codes' semantic relationships and temporal dependencies representing diagnoses, medications, and procedures. Pre-training models on massive EHR datasets can help contextualize data with information not included in smaller task-specific datasets (Rasmy et al., 2021).

Future Research Grinsztajn et al. (2022) note that the underperformance of neural networks on tabular data may stem from a lack of inductive biases—especially when dealing with uninformative or noisy features, which are common in medical data. Future research could explore further the integration of inductive biases using graph or mathematical representations of domain knowledge. For example, Kim et al. (2024a) propose a new pre-training architecture for tabular data using graph representations, enabling improved transfer learning across structured datasets.

Additionally, given the growing interest in medical foundation models, it may be valuable to investigate how pre-training tasks can better exploit fine-grained relationships between clinical codes—potentially improving the quality of learned representations in structured medical data. In addition, though Steinberg et al. (2023) show improved results on pre-trained models compared to trained from scratch, the effect of the pre-training dataset should be studied in more depth. For instance, the impact of the size of the dataset or the distribution shift compared to the downstream task should be assessed. Furthermore, reinforcement learning with human feedback—used, for example, in natural language processing by (Ouyang et al., 2022)—could offer a way to adapt model behavior to clinical expertise, as also explored in other alignment strategies (Yao et al., 2023). This could also be leveraged for tabular datasets.

4.5 Validation

Validation of machine learning models incorporates explainability, generalizability, and bias analysis, which can be grounded in domain knowledge.

A survey by Tonekaboni et al. (2019) highlights that clinicians view *explainability* as a justification tool in clinical workflows. To that end, clinicians must be able to relate model features and outputs to medical reasoning. Explainability methods support clinicians in understanding which features the model considers vital for its decisions (Vimbi et al., 2024).

In addition, auditing frameworks (Drenkow et al., 2025) can enable structured identification of dataset "shortcuts" by comparing feature importance against clinical plausibility. Complementing this, medical literature and clinician insight offer valuable knowledge about known confounders or spurious correlations (Meng et al., 2022).

It is also important to assess model generalizability across patient populations and hospitals. One aspect is to appropriately select metrics and dataset splits. Expert insight can also provide information into possible sources of dataset shift, such as variations in clinical workflows or patient populations. Subbaswamy et al. (2021) propose, for example, a method to evaluate how a model can generalize to shifts in laboratory test ordering.

Finally, it is also crucial to consider the baselines against which machine learning methods will be compared to, as even naive methods can show surprisingly good results. For instance, naive forecasting often shows competitive performance in financial forecasting tasks (Hewamalage et al., 2023). In clinical settings, domain knowledge could be used to construct naive rule-based baselines to validate clinical applications.

Future Research Although current explainability methods increase transparency and trust, they remain approximations of the model's internal logic, can introduce their uncertainties, and may not be suited for clinical decision validation (Ghassemi et al., 2021). Indeed, they cannot guarantee the correctness of predictions or justify their adoption in practice (Ghassemi et al., 2021).

Similarly, while valuable for evaluating model robustness and generalizability, cross-dataset testing assesses performance after distribution shifts have occurred. Future work could prioritize proactive strategies to build more resilient systems that mitigate or validate such shifts in advance, for instance, through synthetic data or causal modeling informed by clinical expertise.

In bias analysis, incorporating structured medical knowledge and recent experimental findings could help identify and address harmful shortcuts. Additionally, synthetic data could be used to generate slightly modified test datasets to assess the robustness of the model to changes that should not be medically relevant to outputs.

5 Discussion

As medical machine learning becomes increasingly prominent, incorporating domain knowledge is vital. Some approaches emphasize the scalability and diversity of large datasets, relying, for instance, on pre-trained models (Steinberg et al., 2023). Others prioritize the structured integration of domain knowledge using ontologies or graphs (Sirocchi et al., 2024). This becomes especially important when dealing with heterogeneous, high-dimensional, or noisy data.

However, access to expert input and curated databases can be limited, and integrating this knowledge effectively is often complex. In addition, clinical practices and medical understanding evolve, and relying on outdated ontologies or prior assumptions may introduce biases. Moreover, models trained on historical data may learn and reinforce prior clinical behaviors, leading to the risk of self-fulfilling prophecies in real-world decision support systems (De-Arteaga and Elmer, 2023). Furthermore, relying too heavily on domain constraints can unintentionally limit the discovery of novel patterns or rare cases. Thus, further empirical evaluations should assess the benefits of knowledge integration methods across medical datasets of different types and quality.

In general, we first recommend early discussions with medical partners to determine potential biases and confounders. While confounders can be unavoidable for retrospective studies, they should be recognized as limitations. Domain knowledge should also be included during data preprocessing to harmonize values following ontologies and guidelines or to assess the reasons for missing data and impute them accordingly. Domain knowledge can also engineer medically relevant features or integrate information from knowledge bases for feature selection. Moreover, model training can leverage pre-trained models or mathematical relationships. Finally, validation should be based on clinical expertise, and potential generalizability should be assessed for other patient populations or hospital settings.

While this process can be time-consuming, recent studies suggest that domain knowledge integration can be automated by leveraging foundation models for knowledge extraction (Kresevic et al., 2024) and its integration in the pipeline (Hollmann et al., 2023b). This paves the way toward scalable medical deep-learning models. Yet, medical foundation models also need to be evaluated in terms of privacy preservation, bias propagation, and generalizability. Recently, studies have led benchmarking efforts for scientific foundation models. Chen et al. (2024) show that while expert knowledge did not always improve code validity, it consistently increased success rates—supporting the idea that domain expertise can improve model outcomes, and its inclusion should be further studied for foundation models. However, medical machine learning on complex tabular datasets cannot rely yet on end-to-end LLMs.

Closer collaboration between the fields of healthcare and tabular machine learning could leverage deep learning advances to design models that integrate domain knowledge more efficiently. Promising research directions include adapting and validating automated approaches for domain knowledge integration and transfer learning for tabular data (Kim et al., 2024a).

6 Limitations

The current study presents several limitations that should be acknowledged. The presented work is not a systematic review and does not aim to cover all relevant literature comprehensively. Thus, it has been influenced by the authors' experiences within the field of medical machine learning.

In addition, while we propose an overview and diverse examples for integrating domain knowledge into the medical machine learning pipeline, we do not offer concrete recommendations that are applicable to all use cases. Indeed, the appropriate approach may vary depending on the medical context and application. Therefore, we encourage interdisciplinary discussions between medical experts and machine learning practitioners to define a concrete guide collaboratively.

Moreover, the efficacy of the discussed methods of domain knowledge integration may vary according to data quality. We do not offer a systematic assessment of these integration methods on various data types, which would be valuable in gaining a deeper understanding of the impact of domain knowledge.

Finally, our focus was limited to tabular data. Integrating domain knowledge into multimodal machine learning models, which utilize data such as text, images, or time series, represents an important direction for future research, but was beyond the scope of this work.

References

Leila Ahmadian, Mariette van Engen-Verheul, Ferishta Bakhshi-Raiez, Niels Peek, Ronald Cornet, and Nicolette F de Keizer. 2011. The role of standardized data and terminological systems in computerized clinical decision support systems: literature review and survey. *International journal of medical informatics*, 80(2):81–93.

Md Manjurul Ahsan, Shahana Akter Luna, and Zahed Siddique. 2022. Machine-learning-based disease diagnosis: A comprehensive review. In *Healthcare*, volume 10, page 541. MDPI.

Giuseppina Albano and Virginia Giorno. 2006. A stochastic model in tumor growth. *Journal of Theoretical Biology*, 242(2):329–336.

Rawan AlSaad, Alaa Abd-Alrazaq, Sabri Boughorbel, Arfan Ahmed, Max-Antoine Renault, Rafat Damseh, and Javaid Sheikh. 2024. Multimodal large language models in health care: applications, challenges, and future outlook. *Journal of medical Internet research*, 26:e59505.

Melanie L Bell, Mallorie Fiero, Nicholas J Horton, and Chiu-Hsieh Hsu. 2014. Handling missing data in rcts; a review of the top medical journals. *BMC medical research methodology*, 14:1–8.

Tommaso Bendinelli, Artur Dox, and Christian Holz. 2025. Exploring llm agents for cleaning tabular machine learning datasets. In *ICLR 2025 Workshop on Foundation Models in the Wild*.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Jennifer Brescoll and Steven Daveluy. 2015. A review of vitamin b12 in dermatology. *American journal of clinical dermatology*, 16:27–33.

M Alan Brookhart, Til Stürmer, Robert J Glynn, Jeremy Rassen, and Sebastian Schneeweiss. 2010. Confounding control in healthcare database research: challenges and potential approaches. *Medical care*, 48(6):S114–S120.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730.

Javier Castell-Díaz, Jose Antonio Miñarro-Giménez, and Catalina Martínez-Costa. 2023. Supporting snomed ct postcoordination with knowledge graph embeddings. *Journal of Biomedical Informatics*, 139:104297.

Eunsuk Chang and Javed Mostafa. 2021. The use of snomed ct, 2013-2020: a literature review. *Journal of the American Medical Informatics Association*, 28(9):2017–2026.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. 2018. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085.

Zekai Chen, Mariann Micsinai Balan, and Kevin Brown. 2023. Language models are few-shot learners for prognostic prediction. *ArXiv*, abs/2302.12692.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, et al. 2024. Scienceagentbench:

Toward rigorous assessment of language agents for data-driven scientific discovery. *arXiv preprint arXiv:2410.05080*.

Iris C Cramer, Eline GM Cox, Jip WTM de Kok, Jacqueline Koeze, Martje Visser, Hjalmar R Bouma, Ashley De Bie Dekker, Iwan CC van der Horst, R Arthur Bouwman, and Bas CT van Bussel. 2025. Quantification of facial cues for acute illness: a systematic scoping review. *Intensive Care Medicine Experimental*, 13(1):17.

Salvatore Cuomo, Vincenzo Schiano Di Cola, Fabio Giampaolo, Gianluigi Rozza, Maziar Raissi, and Francesco Piccialli. 2022. Scientific machine learning through physics–informed neural networks: Where we are and what's next. *Journal of Scientific Computing*, 92(3):88.

Olivier Curé. 2012. Improving the data quality of drug databases using conditional dependencies and ontologies. *Journal of Data and Information Quality (JDIQ)*, 4(1):1–21.

Tirtharaj Dash, Sharad Chitlangia, Aditya Ahuja, and Ashwin Srinivasan. 2022. A review of some techniques for inclusion of domain-knowledge into deep neural networks. *Scientific Reports*, 12(1):1040.

Maria De-Arteaga and Jonathan Elmer. 2023. Self-fulfilling prophecies and machine learning in resuscitation science. *Resuscitation*, 183:109622.

Timo M Deist, Andrew Patti, Zhaoqi Wang, David Krane, Taylor Sorenson, and David Craft. 2019. Simulation-assisted machine learning. *Bioinformatics*, 35(20):4072–4080.

Nathan Drenkow, Mitchell Pavlak, Keith Harrigian, Ayah Zirikly, Adarsh Subbaswamy, and Mathias Unberath. 2025. Detecting dataset bias in medical ai: A generalized and modality-agnostic auditing framework. *arXiv preprint arXiv:2503.09969*.

Issam El Naqa, Aleksandra Karolak, Yi Luo, Les Folio, Ahmad A Tarhini, Dana Rollison, and Katia Parodi. 2023. Translation of ai into oncology clinical practice. *Oncogene*, 42(42):3089–3097.

Christophe Gaudet-Blavignac, Vasiliki Foufi, Mina Bjelogrlic, Christian Lovis, et al. 2021. Use of the systematized nomenclature of medicine clinical terms (snomed ct) for processing free text in health care: systematic scoping review. *Journal of medical Internet research*, 23(1):e24594.

Faisal Ghaffar, Nadine M Furtado, Imad Ali, and Catherine Burns. 2025. Diagnostic decision-making variability between novice and expert optometrists for glaucoma: Comparative analysis to inform ai system design. *JMIR Medical Informatics*, 13:e63109.

Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew Beam. 2021. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet. Digital health*, 3 11:e745–e750.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. 2022. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35:507–520.

Michele Groves, Peter O'Rourke, and Heather Alexander. 2003. The clinical reasoning characteristics of diagnostic experts. *Medical teacher*, 25(3):308–313.

Sebastien Haneuse. 2016. Distinguishing selection bias and confounding bias in comparative effectiveness research. *Medical care*, 54(4):e23–e29.

SM Shamimul Hasan, Donna Rivera, Xiao-Cheng Wu, J Blair Christian, and Georgia Tourassi. 2019. A knowledge graph approach for the secondary use of cancer registry data. In *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 1–4. IEEE.

Ahatsham Hayat and Mohammad Rashedul Hasan. 2024. Claim your data: Enhancing imputation accuracy with contextual large language models. *arXiv preprint arXiv:2405.17712*.

Hansika Hewamalage, Klaus Ackermann, and Christoph Bergmeir. 2023. Forecast evaluation for data scientists: common pitfalls and best practices. *Data Mining and Knowledge Discovery*, 37(2):788–832.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. 2023a. Tabpfn: A transformer that solves small tabular classification problems in a second. In *The Eleventh International Conference on Learning Representations*.

Noah Hollmann, Samuel G. Müller, and Frank Hutter. 2023b. Large language models for automated data science: Introducing caafe for context-aware automated feature engineering. In *Neural Information Processing Systems*.

Madelon Hulsebos, Paul Groth, and Çagatay Demiralp. 2023. Adatyper: Adaptive semantic column type detection. *CoRR*.

KJ Jager, C Zoccali, A Macleod, and FW Dekker. 2008. Confounding: what it is and how to deal with it. *Kidney international*, 73(3):256–260.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Johnny Kahlert, Sigrid Bjerge Gribsholt, Henrik Gammelager, Olaf M Dekkers, and George Luta. 2017. Control of confounding in the analysis phase–an overview for clinicians. *Clinical epidemiology*, pages 195–204.

George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. 2021. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440.

Myung Jun Kim, Léo Grinsztajn, and Gaël Varoquaux. 2024a. Carte: pretraining and transfer for tabular learning. In *Proceedings of the 41st International Conference on Machine Learning*, pages 23843–23866.

Subin Kim, Jungmin Son, Minyoung Jung, and Youngjun Kwak. 2024b. Expertise-centric prompting framework for financial tabular data generation using pre-trained large language models. In *NeurIPS 2024 Third Table Representation Learning Workshop*.

Simone Kresevic, Mauro Giuffrè, Milos Ajcevic, Agostino Accardo, Lory S Crocè, and Dennis L Shung. 2024. Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework. *NPJ digital medicine*, 7(1):102.

Jaris Küken, Lennart Purucker, and Frank Hutter. 2025. Large language models engineer too many simple features for tabular data. In *NeurIPS 2024 Third Table Representation Learning Workshop*.

Florian Leiser, Sascha Rank, Manuel Schmidt-Kraepelin, Scott Thiebes, and Ali Sunyaev. 2023. Medical informed machine learning: A scoping review and future research directions. *Artificial Intelligence in Medicine*, 145:102676.

Dongdong Lin, Ji-Gang Zhang, Jingyao Li, Chao Xu, Hong-Wen Deng, and Yu ping Wang. 2016. An integrative imputation method based on multi-omics datasets. *BMC Bioinformatics*, 17.

Louis Lin and Harvey Wong. 2017. Predicting oral drug absorption: mini review on physiologically-based pharmacokinetic models. *Pharmaceutics*, 9(4):41.

Donald AB Lindberg, Betsy L Humphreys, and Alexa T McCray. 1993. The unified medical language system. *Yearbook of medical informatics*, 2(01):41–51.

Lingchao Mao, Hairong Wang, Leland S Hu, Nhan L Tran, Peter D Canoll, Kristin R Swanson, and Jing Li. 2024. Knowledge-informed machine learning for cancer diagnosis and prognosis: a review. *IEEE Transactions on Automation Science and Engineering*.

Theresa A McDonagh, Marco Metra, Marianna Adamo, Roy S Gardner, Andreas Baumbach, Michael Böhm, Haran Burri, Javed Butler, Jelena Čelutkienė, Ovidiu Chioncel, et al. 2023. 2023 focused update of the 2021 esc guidelines for the diagnosis and treatment of acute and chronic heart failure: developed by the task force for the diagnosis and treatment of acute and chronic heart failure of the european society of cardiology (esc) with the special contribution of the heart failure association (hfa) of the esc. *European heart journal*, 44(37):3627–3639.

Chuizheng Meng, Loc Trinh, Nan Xu, James Enouen, and Yan Liu. 2022. Interpretability and fairness evaluation of deep learning models on mimic-iv dataset. *Scientific Reports*, 12(1):7166.

George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. 2021. Umlsbert: Clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1744–1753.

Lawrence Middleton, Ioannis Melas, Chirag Vasavda, Arwa Bin Raies, Benedek Rozemberczki, Ryan S. Dhindsa, Justin Dhindsa, Blake Weido, Quanli Wang, Andrew R Harper, Gavin Edwards, Slavé Petrovski, and Dimitrios M Vitsios. 2024. Phenome-wide identification of therapeutic genetic targets, leveraging knowledge graphs, graph neural networks, and uk biobank data. *Science Advances*, 10.

Mark D Miller, Craig Steinmaus, Mari S Golub, Rosemary Castorina, Ruwan Thilakartne, Asa Bradman, and Melanie A Marty. 2022. Potential impacts of synthetic food dyes on activity and attention in children: a review of the human and animal evidence. *Environmental Health*, 21(1):45.

Fatemeh Nassajian Mojarrad, Lorenzo Bini, Thomas Matthes, and Stephane Marchand-Maillet. 2024. Injecting hierarchical biological priors into graph neural networks for flow cytometry prediction. In *ICML 2024 Workshop on Efficient and Accessible Foundation Models for Biological Discovery*.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. 2023. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265.

OPhir Nave and Miriam Elbaz. 2021. Artificial immune system features added to breast cancer clinical data for machine learning (ml) applications. *Biosystems*, 202:104341.

Dan Nguyen, Rafe McBeth, Azar Sadeghnejad Barkousaraie, Gyanendra Bohara, Chenyang Shen, Xun Jia, and Steve Jiang. 2020. Incorporating human and learned domain knowledge into training deep neural networks: a differentiable dose-volume histogram and adversarial inspired framework for generating pareto optimal dose distributions in radiation therapy. *Medical physics*, 47(3):837–849.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Richard L Page, José A Joglar, Mary A Caldwell, Hugh Calkins, Jamie B Conti, Barbara J Deal, NA Mark Estes, Michael E Field, Zachary D Goldberger, Stephen C Hammill, et al. 2016. 2015 acc/aha/hrs guideline for the management of adult patients with supraventricular tachycardia: a report of the american college of cardiology/american heart association task force on clinical practice guidelines and the heart rhythm society. *Journal of the American College of Cardiology*, 67(13):e27–e115.

Giulia Perciballi, Federica Granese, Ahmad Fall, Farida ZEHRAOUI, Edi Prifti, and Jean-Daniel Zucker. 2024. Adapting tabPFN for zero-inflated metagenomic data. In *NeurIPS 2024 Third Table Representation Learning Workshop*.

Vasileios C Pezoulas, Dimitrios I Zaridis, Eugenia Mylona, Christos Androutsos, Kosmas Apostolidis, Nikolaos S Tachos, and Dimitrios I Fotiadis. 2024. Synthetic data generation methods in healthcare: A review on open-source tools and methods. *Computational and structural biotechnology journal*.

Julia K Pilowsky, Rosalind Elliott, and Michael A Roche. 2024. Data cleaning for clinician researchers: Application and explanation of a data-quality framework. *Australian Critical Care*, 37(5):827–833.

Paloma Rabaey, Henri Arno, Stefan Heytens, and Thomas Demeester. 2024. Synsum - synthetic benchmark with structured and unstructured medical records. *ArXiv*, abs/2409.08936.

Marsha A Raebel, Susan Shetterly, Christine Y Lu, James Flory, Joshua J Gagne, Frank E Harrell, Kevin Haynes, Lisa J Herrinton, Elisabetta Patorno, Jennifer Popovic, et al. 2016. Methods for using clinical laboratory test results as baseline confounders in multi-site observational database studies when missing data are expected. *Pharmacoepidemiology and drug safety*, 25(7):798–814.

Prabhu Rajagopalan and Marc R. Gastonguay. 2003. Population pharmacokinetics of ciprofloxacin in pediatric patients. *The Journal of Clinical Pharmacology*, 43.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86.

Yucheng Ruan, Xiang Lan, Jingying Ma, Yizhi Dong, Kai He, and Mengling Feng. 2024. Language modeling on tabular data: A survey of foundations, techniques and evolution. *CoRR*.

Camilo Ruiz, Hongyu Ren, Kexin Huang, and Jure Leskovec. 2023. High dimensional, tabular deep learning with an auxiliary knowledge graph. *Advances in Neural Information Processing Systems*, 36:26348–26371.

Elizaveta Savchenko, Ariel Rosenfeld, and Svetlana Bunimovich-Mendrazitsky. 2023. Mathematical modeling of bcg-based bladder cancer treatment using socio-demographics. *Scientific Reports*, 13(1):18754.

FP Schmidt, NJ Glaser, O Schreiner, T Münzel, and M Weber. 2015. „do not rescuscitate “–auswirkungen der einführung eines standardisierten formulars auf therapiebegrenzungen in der klinischen praxis. *DMW-Deutsche Medizinische Wochenschrift*, 140(15):e159–e165.

Sebastian C Semler, Frank Wissing, and Ralf Heyder. 2018. German medical informatics initiative. *Methods of information in medicine*, 57(S 01):e50–e56.

Jingpu Shi, Dong Wang, Gino Tesei, and Beau Norgeot. 2022. Generating high-fidelity privacy-conscious synthetic patient data for causal effect estimation with multiple treatments. *Frontiers in Artificial Intelligence*, 5:918813.

Xi Shi, Charlotte Prins, Gijs Van Pottelbergh, Pavlos Mamouris, Bert Vaes, and Bart De Moor. 2021. An automated data cleaning method for electronic health records by incorporating clinical knowledge. *BMC Medical Informatics and Decision Making*, 21:1–10.

Christel Sirocchi, Alessandro Bogliolo, and Sara Montagna. 2024. Medical-informed machine learning: integrating prior knowledge into medical decision systems. *BMC Medical Informatics and Decision Making*, 24(Suppl 4):186.

Dylan Slack and Sameer Singh. 2023. Tablet: Learning from instructions for tabular data. *ArXiv*, abs/2304.13188.

Ethan Steinberg, Jason Alan Fries, Yizhe Xu, and Nigam Shah. 2023. Motor: A time-to-event foundation model for structured medical records. In *The Twelfth International Conference on Learning Representations*.

Adarsh Subbaswamy, Roy Adams, and Suchi Saria. 2021. Evaluating model robustness and stability to dataset shift. In *International conference on artificial intelligence and statistics*, pages 2611–2619. PMLR.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Peakman, and Rory Collins. 2015. UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, 12(3):e1001779.

Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. 2024. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 645–654.

Kexuan Sun, Harsha Rayudu, and Jay Pujara. 2021. A hybrid probabilistic approach for table understanding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 4366–4374.

Mohammad Tabatabai, David Keith Williams, and Zoran Bursac. 2005. Hyperbolastic growth models: theory and application. *Theoretical Biology and Medical Modelling*, 2:1–13.

Zoë Tieges, Jonathan J Evans, Karin J Neufeld, and Alasdair MJ MacLullich. 2018. The neuropsychology of delirium: advancing the science of delirium assessment. *International journal of geriatric psychiatry*, 33(11):1501–1511.

Nenad Tomašev, Natalie Harris, Sebastien Baur, Anne Mottram, Xavier Glorot, Jack W Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Valerio Magliulo, et al. 2021. Use of deep learning to develop continuous-risk models for adverse event prediction from electronic health records. *Nature Protocols*, 16(6):2765–2787.

Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. 2019. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*, pages 359–380. PMLR.

Allan Tucker, Zhenchen Wang, Ylenia Rotalinti, and Puja Myles. 2020. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *NPJ digital medicine*, 3(1):147.

Viswan Vimbi, Noushath Shaffi, and Mufti Mahmud. 2024. Interpreting artificial intelligence models: a systematic review on the application of lime and shap in alzheimer's disease detection. *Brain Informatics*, 11(1):10.

Stein Emil Vollset, Hazim S Ababneh, Yohannes Habtegiorgis Abate, Cristiana Abbafati, Rouzbeh Abbasgholizadeh, Mohammadreza Abbasian, Hedayat Abbastabar, Abdallah HA Abd Al Magied, Samar Abd ElHafeez, Atef Abdelkader, et al. 2024. Burden of disease scenarios for 204 countries and territories, 2022–2050: a forecasting analysis for the global burden of disease study 2021. *The Lancet*, 403(10440):2204–2256.

Laura Von Rueden, Sebastian Mayer, Katharina Beckh, Bogdan Georgiev, Sven Giesselbach, Raoul Heese, Birgit Kirsch, Julius Pfrommer, Annika Pick, Rajkumar Ramamurthy, et al. 2021. Informed machine learning–a taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):614–633.

Jared Willard, Xiaowei Jia, Shaoming Xu, Michael Steinbach, and Vipin Kumar. 2022. Integrating scientific knowledge with machine learning for engineering and environmental systems. *ACM Computing Surveys*, 55(4):1–37.

Katarzyna Woźnica, Piotr Wilczyński, and Przemysław Biecek. 2024. Sefnet: Linking tabular datasets with semantic feature nets. *Available at SSRN 4811308*.

Xingyu Wu, Zhenchao Tao, Bingbing Jiang, Tianhao Wu, Xin Wang, and Huanhuan Chen. 2022. Domain knowledge-enhanced variable selection for biomedical data analysis. *Information Sciences*, 606:469–488.

Magdalena Wysocka, Oskar Wysocki, Marie Zufferey, Dónal Landers, and André Freitas. 2023. A systematic review of biologically-informed deep learning models for cancer: fundamental trends for encoding and interpreting oncology data. *BMC bioinformatics*, 24(1):198.

Xiaozheng Xie, Jianwei Niu, Xuefeng Liu, Zhengsu Chen, Shaojie Tang, and Shui Yu. 2021. A survey on incorporating domain knowledge into deep learning for medical image analysis. *Medical Image Analysis*, 69:101985.

Zonghai Yao, Benjamin J Schloss, and Sai P. Selvaraj. 2023. Improving summarization with human edits. In *Conference on Empirical Methods in Natural Language Processing*.

Jim Young, Patrick Graham, and Richard Penny. 2009. Using bayesian networks to create synthetic data. *Journal of Official Statistics*, 25(4):549–567.

Fuad Abu Zahra and Rohit J Kate. 2024. Obtaining clinical term embeddings from snomed ct ontology. *Journal of Biomedical Informatics*, 149:104560.

Dan Zhang, Yoshihiko Suhara, Jinfeng Li, Madelon Hulsebos, Cagatay Demiralp, and Wang-Chiew Tan. 2020. Sato: Contextual semantic type detection in tables. *Proceedings of the VLDB Endowment*, 13(11).

Tianping Zhang, Shaowen Wang, Shuicheng Yan, Li Jian, and Qian Liu. 2023. Generative table pre-training empowers models for tabular prediction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14836–14854.