



Review

# Considerations for the Use of Machine Learning Extracted Real-World Data to Support Evidence Generation: A Research-Centric Evaluation Framework

Melissa Estevez <sup>1</sup>, Corey M. Benedum <sup>1</sup>, Chengsheng Jiang <sup>1</sup>, Aaron B. Cohen <sup>1,2</sup>, Sharang Phadke <sup>1</sup>, Somnath Sarkar <sup>1</sup> and Selen Bozkurt <sup>1,\*</sup>

<sup>1</sup> Flatiron Health, Inc., 233 Spring Street, New York, NY 10013, USA; mhedberg@flatiron.com (M.E.); corey.benedum@flatiron.com (C.M.B.); chengsheng.jiang@flatiron.com (C.J.); acohen@flatiron.com (A.B.C.); sharang.phadke@gmail.com (S.P.); ssarkar@flatiron.com (S.S.)

<sup>2</sup> Department of Medicine, NYU Grossman School of Medicine, New York, NY 10016, USA

\* Correspondence: selen.bozkurt@flatiron.com

**Simple Summary:** Many patient clinical characteristics, such as diagnosis dates, biomarker status, and therapies received, are only available as unstructured text in electronic health records. Obtaining this information for research purposes is a difficult and costly process, requiring trained clinical experts to manually review patient documents. Machine Learning techniques offer a promising solution for efficiently extracting clinically relevant information from unstructured text found in patient documents. However, the use of data produced with machine learning techniques for research purposes introduces unique challenges in assessing validity and generalizability to different cohorts of interest. To enable the effective and accurate use of such data for research purposes, we developed an evaluation framework to be utilized by model developers, data users, and other stakeholders. This framework can serve as a baseline to contextualize the quality, strengths, and limitations of using data produced with machine learning techniques for research purposes.

**Abstract:** A vast amount of real-world data, such as pathology reports and clinical notes, are captured as unstructured text in electronic health records (EHRs). However, this information is both difficult and costly to extract through human abstraction, especially when scaling to large datasets is needed. Fortunately, Natural Language Processing (NLP) and Machine Learning (ML) techniques provide promising solutions for a variety of information extraction tasks such as identifying a group of patients who have a specific diagnosis, share common characteristics, or show progression of a disease. However, using these ML-extracted data for research still introduces unique challenges in assessing validity and generalizability to different cohorts of interest. In order to enable effective and accurate use of ML-extracted real-world data (RWD) to support research and real-world evidence generation, we propose a research-centric evaluation framework for model developers, ML-extracted data users and other RWD stakeholders. This framework covers the fundamentals of evaluating RWD produced using ML methods to maximize the use of EHR data for research purposes.

**Keywords:** artificial intelligence; deep learning; machine learning; oncology; personalized medicine



**Citation:** Estevez, M.; Benedum, C.M.; Jiang, C.; Cohen, A.B.; Phadke, S.; Sarkar, S.; Bozkurt, S. Considerations for the Use of Machine Learning Extracted Real-World Data to Support Evidence Generation: A Research-Centric Evaluation Framework. *Cancers* **2022**, *14*, 3063. <https://doi.org/10.3390/cancers14133063>

Academic Editors: Andreas Stadlbauer, Anke Meyer-Baese and Max Zimmermann

Received: 25 May 2022

Accepted: 17 June 2022

Published: 22 June 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Real-world data (RWD) leveraged from electronic health records (EHRs) can provide valuable insights into patients' treatment pathways and facilitate health outcomes research. Specifically in oncology, where the treatment landscape is constantly evolving to include numerous targeted therapies for often rare populations of patients, RWD and real-world evidence (RWE) play a pivotal role in supplementing clinical trials [1–3] and the use of RWD to help clinicians, researchers, and other stakeholders in the healthcare ecosystem understand the ever-changing oncology treatment landscape is important. While some

data elements are captured and stored in a structured format (e.g., lab results), the majority of data—including critical elements such as diagnosis dates, genomic test results, and adverse events—are captured and stored in unstructured (e.g., pathology reports, clinical narratives, etc.) formats [4]. Collecting these data from unstructured text has largely relied on manual chart review by an expert clinical abstractor, which can be challenging and time-intensive. Natural language processing (NLP) techniques address this challenge by providing automated solutions for processing free-text clinical notes to extract task-specific information, such as the metastasis status of a patient or adverse event to a treatment.

Today most of the NLP approaches use machine learning (ML) methods and have been applied across many clinical domains for a variety of information extraction tasks such as identifying a group of patients who have a specific diagnosis, share common characteristics, or show a progression of the disease including breast cancer, colorectal cancer, prostate cancer, and tuberculosis, etc. [3,5–11]. With the promises ML brings to RWD generation, in terms of scalability via reduced manual chart review burden [12–14], researchers and data vendors alike have already adopted ML to extract information documented in the EHR (ML-extracted RWD). Despite the increasing use of ML approaches for clinical information extraction, using these methods for research still introduces unique challenges, not only for specific tasks but also in terms of generalizability to different sub-cohorts.

Several interdisciplinary teams have proposed guidelines for developing and evaluating ML applications in healthcare [15–20]. These guidelines primarily focus on ML models that are intended to inform clinical decision-making at the point of care and often involve models that classify patients into different characteristics or predict future events (e.g., predicting a patient’s risk level or likelihood for a future event). As a result, these guidelines do not account for the unique considerations necessary for applying the use of ML-extracted data, collected from unstructured text in the EHRs, to health care outcomes research. Recent FDA draft guidance to assess RWD includes a section about using ML/AI for “unstructured data” in EHRs and encourages transparent reporting of the model and validation details [21]. However, the guidance does not include recommendations pertaining to the use of ML-extracted RWD from EHR for research purposes to generate RWE. The success of using ML-extracted RWD for research purposes (e.g., informing clinical development/trial design, regulatory decision making, etc.) broadly depends on its reliability, generalizability to cohorts of interest, and its fairness across demographic subgroups. Therefore, it is crucial to assess the performance of ML-extracted RWD for relevant clinical and demographic aspects to avoid harmful consequences which might systematically exclude some groups of patients [22]. Sendak et al. suggested including a “Warning” section for all clinical ML models developed for point of care [18] an approach that can be extended to ML-extracted RWD.

In order to enable successful, responsible, and effective use of ML-extracted RWD to support research for evidence generation, we created an evaluation framework for model developers, researchers using ML-extracted variables, and other stakeholders to address these unique considerations. This framework is intended only for RWD that is automatically extracted using ML methods from unstructured EHR documents [8–10] and examines model performance (e.g., sensitivity, PPV, and accuracy) as compared to chart abstraction (data abstracted by clinical experts) to the extent that human labeling is seen as a gold-standard. The goal of this framework is to cover the fundamentals of evaluating the performance, limitations, and fit-for-purpose use of ML-extracted RWD which can be used by different stakeholders for a variety of research purposes within the oncology landscape (e.g., breast and colorectal) as well as other clinical areas (e.g., tuberculosis, COVID-19) [3,5–7].

## **2. A Research-Centric Evaluation Framework for ML-Extracted RWD to Support Research and Evidence Generation**

The research-centric evaluation framework consists of four modular components as shown in Figure 1. While each is an important component for assessing the appropriateness

of the extracted RWD for analysis, the importance might vary based on the use case. This framework assumes that the model prototyping and validation steps are completed and documented using reporting best practices [19]. The intention of this evaluation framework is to be used as the first step before using ML-extracted RWD for research purposes.



**Figure 1.** Evaluation Framework.

### 2.1. Test Set

All components of the evaluation framework should be performed using a well-curated, representative dataset of an appropriate size to assess the generalizability of sub-cohorts derived from the population of interest [23–28]. This dataset, which is referred to as the “test set” here, should not be used during the model development (training and internal validation) process.

### 2.2. Overall Performance Assessment

Overall performance metrics (Table 1) aim to provide a high-level understanding of a given ML-extracted variable’s performance on the held-out test set described above. Performance metrics should help users make decisions based on the quality and the fit-for-use of an ML-extracted variable for common research purposes, such as cohort selection or retrospective analyses. It is also important to consider the clinical perspective while calculating and using these metrics. For example, while evaluating date variables, defining an error window in too narrow a way (example: 3 days) may not be clinically warranted for the use case and underestimate the performance of the date extraction model. Likewise using too large of an error window (example: 90 days) might be less clinically meaningful or trustworthy. This illustrates the importance of balancing performance metrics with the clinical utility to align on an error window that is appropriate for the specific use case.

### 2.3. Stratified Performance

Stratified performance analysis indicates whether model performance differs among certain subgroups (e.g., whether algorithmic bias exists) [29]. Stratified performance analysis can help indicate:

1. Whether model performance may be worse for a particular sub-cohort of interest given analyses often only cover specific sub-populations. For example, if recently diagnosed patients are of interest for an analysis, stratification by initial diagnosis year group can help detect whether there are any changes in documentation patterns (e.g., data shift) over time that is impacting model performance in the target population.
2. Whether model performance may be worse for a demographic group. High model performance does not preclude model errors from being concentrated in specific demographic groups due to bias introduced during model development or chance alone. Evaluating performance stratified by demographic subgroups, such as race and gender, may help to minimize unintentional discrimination caused by the model and ensure model fairness and generalizability. Readers should keep in mind that

these assessments should also be performed during the validation process to ensure all actions needed to develop fair models were taken into account and at this stage, only the final findings are reported as confirmation or data use limitations.

- Whether the performance is differential with respect to important covariates, such as treatment status, stage, biomarker status, and demographic characteristics, as many analytic use cases are focused on sub-cohorts that are receiving specific treatments. Differential error makes it harder to predict if and how model errors lead to bias in an analysis in which these variables are used as covariates.

**Table 1.** Example performance metrics.

Variable Type	Example ML-Extracted Variable	Example Performance Metric
Categorical	Diagnosis (yes/no)	Sensitivity Positive predictive value (PPV or precision) Specificity Negative predictive value (NPV) Accuracy Predicted prevalence vs. abstracted prevalence Calibration plots (if applicable)
Date	Diagnosis date	Sensitivity with a $\pm$ n-day window PPV with a $\pm$ n-day window <sup>1</sup> Distribution of date errors
Continuous	Lab value	Sensitivity, PPV, and accuracy for classifying the result as within vs. outside the normal range Sensitivity, PPV, and accuracy for classifying the result within $\pm X$ of the true value Mean absolute error (MAE)

<sup>1</sup>: The proportion of patients' human-abstracted as having the diagnosis that is also correctly identified as having the diagnosis by the model and where the ML-extracted diagnosis date is within  $\pm n$  days of the abstracted diagnosis date or both abstracted and ML-extracted dates are unknown.

Table 2 provides examples of stratifying variables that may be relevant for a cancer dataset and may not be applicable to other fields.

**Table 2.** Stratified analysis steps and example variables.

Goal	Example Strata Variables
(I.) Understand performance in sub-cohorts of interest	Year of diagnosis (e.g., before vs. after year x) Treatment status (treated vs. not treated) Biomarker status (positive vs. negative)
(II.) Fairness	Race and ethnicity group Gender Age group Insurance status
(III.) Risk for statistical bias in analysis	Treatment setting (Academic vs. Community) Cancer stage at diagnosis Age at diagnosis Cancer histology Smoking status Treatment status (treated vs. not treated) Biomarker status (positive vs. negative)

#### 2.4. Quantitative Error Analysis

Quantitative error analysis refers to the comparison of misclassified patients to correctly classified patients in terms of their demographic and clinical characteristics or outcomes. For example, let us assume there is a binary ML-extracted variable that is used to

select a cohort of patients (e.g., we are interested in patients with characteristic A, where characteristic A is defined by an ML-extracted variable). Using the ML-extracted variable to define our study cohort, we select both true positives and false positives as the subjects for our analysis and incorrectly exclude false negatives (Figure 2) potentially biasing downstream analyses.

	<b>Abstracted as having characteristic A</b>	<b>Abstracted as not having characteristic A</b>	
<b>ML-extracted as having characteristic A</b>	<b>True positives</b> <i>patients with characteristic A that were correctly selected by the ML-extracted variable</i>	<b>False positives</b> <i>patients that do not have characteristic A that were falsely selected by the ML-extracted variable</i>	<b>ML-extracted study cohort</b>
<b>ML-extracted as not having characteristic A</b>	<b>False negatives</b> <i>patients with characteristic A that were falsely excluded by the ML-extracted variable</i>	<b>True negatives</b> <i>patients without characteristic A that were correctly excluded by the ML-extracted variable</i>	
	<b>Abstracted study cohort</b>		

Figure 2. Confusion matrix for model errors.

Through error analysis, we can learn how model errors impact what we observe in our selected study cohort and the potential biases introduced into downstream analyses by making the following comparisons in Table 3.

Table 3. Examples for comparison of errors and their interpretation.

Comparison	Usefulness/Interpretation
True positives vs. False negatives	<p>This comparison informs whether patients incorrectly excluded from the study cohort differ from those correctly included with respect to patient characteristics or outcomes.</p> <p>If true positive and false negative patients appear similar:</p> <ul style="list-style-type: none"> <li>• Model misclassification may be random and excluded patients will most likely have a minimal impact on analysis results.</li> </ul> <p>If true positive and false negative patients appear different:</p> <ul style="list-style-type: none"> <li>• Model misclassification may be systematic and excluded patients may impact analysis results.</li> </ul>
True positives vs. False positives	<p>This comparison informs whether patients incorrectly included in the study cohort differ from those correctly included with respect to patient characteristics or outcomes.</p> <p>If true positive and false positive patients appear similar:</p> <ul style="list-style-type: none"> <li>• Model misclassification may be random and incorrectly included patients may have a minimal impact on analysis results.</li> </ul> <p>If true positive and false positive patients appear different:</p> <ul style="list-style-type: none"> <li>• Model misclassification may be systematic and incorrectly including patients may impact analysis results.</li> </ul>

For each comparison mentioned above, we can look at a wide range of analyses. For example, we can compare demographic and clinical characteristics of patients correctly classified by the ML-extracted variable vs. misclassified. Similar to stratified performance metrics, this analysis can inform about a model's fairness for intended use cases. We can also look directly at outcomes that are relevant for analyses of interest. In cancer research, common outcomes could include prevalence estimates, treatment patterns, and time-to-event analyses.

### 2.5. Replication of Analytic Use Cases

The goal of replication analyses is to provide insights into how the ML-extracted data perform on a breadth of use cases as compared to the corresponding abstracted variables. Replication analyses can range in complexity, from comparing cohorts (e.g., baseline characteristics and outcomes) selected using the ML-extracted variable vs. abstracted counterpart, to replicating analyses with multiple inclusion/exclusion criteria and complex statistical methodologies (e.g., comparative effectiveness studies or using ML-extracted data as a covariate or outcome variable). While it would not be practical to target an exact use case or all possible use cases, researchers should consider selecting a representative suite of analyses to enable stakeholders to evaluate whether a ML-extracted variable is fit for use for all anticipated use cases.

Model performance metrics such as sensitivity and PPV might not be informative enough to describe how analytical outcomes (such as real-world overall survival (rwOS) estimates or hazard ratios) may be impacted by model errors, as model errors may not be non-differential with respect to the outcome of interest and other relevant covariates. Moreover, conventional performance metrics are unable to describe how model errors interact with each other across multiple ML-extracted variables. Replication of analytic use cases can help contextualize the quality of model-extracted variables for end-use cases in several ways:

1. Replication of analytic use cases allows us to focus on specific sub-cohorts of interest.
2. The impact of model errors may differ based on how the variable is used. For example, a categorical variable can be used to select or stratify the study cohort, or it can be used as a variable in the analysis itself (e.g., as a covariate or in a propensity matching algorithm). A date variable can be used as the index date in a time-to-event analysis (e.g., rwOS from metastatic diagnosis date) or to select the study cohort (e.g., patients that started a particular therapy after metastatic diagnosis).
3. Model errors may be correlated rather than randomly distributed across the patient population. Replication analyses can shed light on the combined model performance which may be higher or lower for a selected cohort.
4. For event-level (rather than patient-level) variables, such as biomarker testing events, model performance metrics relevant to the actual use case can be difficult to define and interpret. For example, a biomarker model may over-predict the number of biomarker tests patients receive, resulting in lower test-level precision metrics. However if the majority of use cases are only interested in the biomarker test result closest to diagnosis, additional false-positive predictions at other temporal points would not introduce bias to the analysis results as long as the patient's biomarker status at diagnosis is correctly predicted.

When evaluating ML-extracted variables through replication analysis, consideration should be given to the data available. If ML-extracted data are available without corresponding abstracted data for a cohort of interest, the analysis may be limited to replicating results published in the literature. While this comparison can indicate external validity, it will not be able to yield any insight as to how model errors may influence the reproduced result. For example, the published and reproduced result may differ despite high model performance simply because of differences in the study population. Conversely, if abstracted and ML-extracted data are available for the cohort of interest, researchers can design *in silico* experiments to understand how model errors may influence the parameter

of interest. To evaluate the potential impact of model errors, sensitivity analyses, such as quantitative bias analysis, can be conducted by researchers to better understand whether model misclassification is systematically introducing bias [30].

### 3. Illustrative Use Case

A model that extracts metastatic diagnosis from the EHR will be used as an illustrative use case to demonstrate how the framework can be applied to a specific ML-extracted variable (Table 4). The ultimate purpose of the use case is to use the ML-extracted metastasis variable to select a study cohort for a health outcomes research question.

**Table 4.** Evaluation framework template for the illustrative example.

Variable: Metastatic Diagnosis (yes/no)		
Model Description <sup>1</sup>		
Inputs to the model include unstructured documents from the EHR (e.g., visit notes, pathology/radiology reports). The output of the model is a binary prediction (yes/no) for whether the patient has a metastatic diagnosis at any time in the record.		
Target Dataset/Population		
The model is used in a dataset that contains patients with non-small cell lung cancer (NSCLC).		
Common Analytic Use Case		
<ul style="list-style-type: none"> <li>• Selecting a cohort of patients who have (or do not have) metastatic disease</li> <li>• Using metastatic status as a covariate or stratifying variable in an analysis</li> </ul>		
ML-Extracted Variable Evaluation		
Components	Description	Hypothetical Results and Findings
Test Set	The size of the test set is selected to achieve a target margin of error for the primary evaluation metric (e.g., sensitivity or PPV) within the minority class (metastatic disease). To measure model performance, a random sample of patients is taken from a NSCLC cohort and withheld from model development.	Patients selected from the target population which is not included in model development
Overall Performance	As the primary use of this variable is to select a cohort of metastatic patients, sensitivity, PPV, specificity, and NPV are measured. To evaluate how well this variable selects a metastatic cohort, emphasis is placed on sensitivity and PPV to understand the proportion of patients missed and the proportion of patients incorrectly included in the final cohort.	Sensitivity <sup>2</sup> = 0.94 PPV <sup>3</sup> = 0.91 Specificity <sup>4</sup> = 0.90 NPV <sup>5</sup> = 0.90
Stratified Performance	Sensitivity and PPV for both Metastatic and Non-metastatic classes are calculated across strata of variables of interest. Stratifying variables are selected with the following goals in mind: <ol style="list-style-type: none"> <li>1. Performance in sub-cohorts of interest (e.g., year of diagnosis)</li> <li>2. Fairness (e.g., race and ethnicity)</li> <li>3. Risk for statistical bias in analysis (e.g., cancer stage at diagnosis)</li> </ol>	Example finding for race and ethnicity: <ul style="list-style-type: none"> <li>• Sensitivity for the “metastatic” class is 5% better for “Black or African American” race group vs. “White”.</li> <li>• PPV for the “metastatic” class is 5% lower for “Black or African American” race group vs. “White”</li> </ul>

Table 4. Cont.

Quantitative Error Analysis	<p>To understand the impact of model errors on the selected study cohort, baseline characteristics and rwOS are evaluated for the following groups</p> <ul style="list-style-type: none"> <li>• True positives vs. false negatives</li> <li>• True positives vs. false positives</li> </ul> <p>Typically, patients with non-metastatic disease have longer survival times than patients with metastatic disease. If model misclassification is random, the inclusion of false positives in the study cohort will result in longer observed survival times. However, if model misclassification is systematic and false positives have survival similar to patients with metastatic disease, then the distribution of survival times may remain relatively unchanged.</p>	<p>Example findings from rwOS analysis *:</p> <ul style="list-style-type: none"> <li>• rwOS** for False Positives (21 months) was similar to True Positives (17 months).</li> </ul> <p>Example findings from baseline characteristic analysis:</p> <ul style="list-style-type: none"> <li>• Compared to true negatives, false positives are less likely to have a history of smoking (86% vs. 91%).</li> </ul>
Replication of Use Cases	<p>Evaluate rwOS from metastatic diagnosis date for patients selected as metastatic by the ML-extracted variable vs. abstracted counterpart (outcomes in the general population)</p>	<p>rwOS for ML extracted cohort: 9.8 months (95% CI 8.92–10.75) rwOS for abstracted cohort: 9.8 months (95% CI 8.92–10.69)</p>

<sup>1</sup>: Model is constructed using snippets of text around key terms related to “metastasis,” and processed by a long short-term memory (LSTM) network to produce a compact vector representation of each sentence. These representations were then processed by additional network layers to produce a final metastatic status prediction [31]. <sup>2</sup>: Sensitivity refers to the proportion of patients abstracted as having a value of a variable (e.g., metastasis = true) that are also ML-extracted as having the same value. <sup>3</sup>: PPV refers to the proportion of patients ML-extracted as having a value of a variable (e.g., metastasis = true) that is also abstracted as having the same value. <sup>4</sup>: Specificity refers to the proportion of patients abstracted as not having a value of a variable (e.g., metastasis = false) that are also ML-extracted as not having the same value. <sup>5</sup>: NPV refers to the proportion of patients ML-extracted as not having a value of a variable (e.g., metastasis = false) that are also abstracted as not having the same value. \*: rwOS analysis was performed using Kaplan–Meier method [32]. \*\*: The index date selected for rwOS calculation can be changed based on the study goals. However, the index date that is selected should be available for all patients, regardless of the concordance of their abstracted and predicted value. In this illustrative example, we provided the rwOS strictly as an example and do not specify the index date as index date selection will be case-dependent.

#### 4. Discussion

While ML has tremendous promise for unlocking the power of RWD it is important to note these models are not always perfect and should be evaluated and monitored for any potential bias or findings to prevent negative consequences on the integrity of the data. For example, systematic misclassification errors for metastatic patients might cause the model to select a skewed cohort of patients and cause unrepresentative results. Therefore, it is critical to use a rigorous evaluation framework that assesses clinical ML applications early and at various stages of their development. However, currently available evaluation and reporting frameworks fall short in assessing ML-extracted RWD, despite growing and potential use cases of these data for research or regulatory decision-making [15–19]. Our goal was to put forth a framework for minimum evaluation standards necessary to understand the quality of ML-extracted RWD for target populations and to expose hidden biases of an information extraction tool for researchers. We believe that before using any ML-generated RWD for research or decision-making purposes, researchers must first check the strengths and limitations of these data sets through a standardized evaluation framework [33].

In addition to our proposed evaluation framework, we provided an example use case for a hypothetical metastatic variable. This use case allowed us to demonstrate the strengths and limitations of this variable which users can take into account while using the ML-extracted variable in their research studies. First, the metastatic variable had over 90% sensitivity and specificity. Second, while sensitivity was 5% higher for “Black or African American” race group vs. “White”, the PPV for the “metastatic” class was 5% lower, which can lead to a ML-selected metastatic cohort having a higher rate of Black or African American patients with false-positive metastatic classification relative to White patients. Third, the errors did not cause a difference in the example outcome analysis



(rwOS). Fourth, the model had sufficient performance to demonstrate similar results with abstracted data when this variable was used to select a study-specific cohort from the target population.

Information extraction using ML enriches the potential utilization of EHR-derived unstructured text for quality improvement, clinical and translational research, and/or regulatory decision-making. Unsurprisingly the importance of automated information extraction applications has been growing throughout the RWD life cycle. While there are many suggestions and guidelines by scientists and regulators to assess the quality and fit-for-purpose of RWD [21,34,35], there is still a gap in evaluating ML-extracted RWD in terms of accuracy, bias, fairness, and error distributions. Moreover, while the use of ML and NLP can improve the power and coverage of RWD, they do not replace the need for a well-thought-out research question and longitudinal data sets.

For most analyses using RWD, multiple ML-extracted variables will be used in cohort selection or be statistically summarized. This “layering” of variables may result in end-to-end variable performance differing from the variable level performance assessment in the test set. For example, if a user is selecting a cohort of patients with ALK+ disease treated with Drug A, patients who are incorrectly classified as receiving Drug A (i.e., false positive) but correctly classified as not having ALK+ disease (i.e., true negative) will still be correctly excluded from this cohort. Thus, when we consider the precision of Drug A within this cohort, it would be higher than the precision measured within the performance assessment framework. While it is not possible to assess all use cases, replicating a set of common use cases including several ML-extracted variables might provide a better fit-for-purpose assessment of the data set created.

The evaluation framework we propose for ML-extracted variables is not without limitations. First, it does not include any thresholds of what level of performance is “good enough”. Hernandez-Boussard et al. (2019) [36] proposed a threshold for regulatory grade extracted data as recall >85% and precision >90%, however, this suggestion does not rely on an empirical study, and the authors aimed to propose a starting benchmark to initiate discussion. Since the minimum acceptable performance threshold may be use-case dependent and there is no commonly accepted definition of “good enough,” we believe the “replication of use case” component of our framework would provide the insights necessary for determining such thresholds for different research scenarios. Our framework can also be combined with other frameworks to assess Regulatory-Grade Data Quality [37] for regulatory use cases.

Second, the stratified performance assessment component of the framework suggests the evaluation of several variables of interest which might become challenging when deciding what action to take once all comparisons are completed. We encourage the users of the framework to select the subset of variables based on their use case and set acceptable error rates in advance of the evaluation process. While choosing the set of variables for this step, it would be beneficial to keep in mind the target cohort, clinical question, and seven sources of harm in ML [38].

Third, our framework does not include continuous model monitoring with data-in-the-wild (i.e., data that are currently unseen but could be encountered in the EHR in the future). For example, a new site could be added to the network of oncology clinics that was not there when the model was trained; thus, continuous monitoring of such models is an important piece to consider for ML-extracted data in recurring datasets. Fourth, our framework does not include investigations regarding the explainability of the models used [5], however, this framework can be extended to account for any advancement in this active field of research.

Finally, we developed and applied this evaluation framework to RWD harnessed for oncology use cases. However, we feel the components of this framework are important to understand when using models to extract variables, irrespective of the disease of interest, and should not be limited or restricted to oncology.

## 5. Conclusions

Innovations in machine learning offer new opportunities and solutions to accelerate health research, including the generation of real-world data at a scale previously unachievable through conventional manual chart reviews. This may provide immense value as demand increases for research into niche subpopulations that require larger RWD cohort sizes. While holding tremendous promise, machine learning solutions introduce unique challenges, such as access to high-quality data for model development, generalizability to cohorts of interest, and lack of model interpretability. To enable the widespread adoption of such methods, machine learning practitioners must describe the methods used to evaluate the model and the implication of model errors in downstream analyses. To this end, we proposed an evaluation framework for research-centric assessment of ML-extracted RWD before the data are used for research purposes. Our evaluation framework provides a structured approach to documenting the strengths, limitations, and applications of ML-extracted RWD. By going beyond standard machine learning metrics with the inclusion of stratified performance analysis, quantitative error analysis, and replication analysis, this framework enables a deeper understanding of potential model biases and how the ML-extracted data may perform in research use cases as compared to the corresponding abstracted variables. Despite its limitations, we feel that our proposed framework can be used as a baseline and with further testing, can be applied to additional use cases and disease types. Ultimately, our framework can be used to define a minimum quality threshold when using ML-extracted RWD for different use cases. As ML-extraction methods continue to advance, this framework can also be extended or modified to accommodate new variable types or uses of ML-extracted variables. We believe the adoption of such frameworks will improve transparency with respect to the quality of ML-extracted variables and can promote the responsible and effective use of ML-extracted RWD.

**Author Contributions:** Conceptualization, M.E., A.B.C., C.J., S.P., S.B.; Methodology, M.E., C.M.B., C.J.; Writing—Original Draft Preparation, M.E., C.M.B., S.B.; Writing—Review & Editing, A.B.C., S.S., S.P.; Supervision, S.B., S.S.; Funding Acquisition, N/A. All authors have read and agreed to the published version of the manuscript.

**Funding:** This study was sponsored by Flatiron Health, which is an independent subsidiary of the Roche Group.

**Institutional Review Board Statement:** The data used in this study were simulated, therefore no IRB approval was necessary.

**Informed Consent Statement:** The data used in this study were simulated, therefore no patient consent was obtained or required.

**Data Availability Statement:** The simulated data used in this study have been originated by Flatiron Health, Inc. These simulated de-identified data may be made available upon request and are subject to a license agreement with Flatiron Health; interested researchers should contact <DataAccess@flatiron.com> to determine licensing terms.

**Acknowledgments:** The authors would like to acknowledge the writing and editing support provided by Hannah Gilham and Cody Patton from Flatiron Health.

**Conflicts of Interest:** At the time of the study, all authors reported employment in Flatiron Health, Inc., an independent subsidiary of Roche, and reported stock ownership in Roche. S.S., S.P., A.B.C., and M.E. also reported equity ownership in Flatiron Health, Inc. C.M.B. also reported stock ownership in Pfizer.

## References

1. Booth, C.M.; Karim, S.; Mackillop, W.J. Real-World Data: Towards Achieving the Achievable in Cancer Care. *Nat. Rev. Clin. Oncol.* **2019**, *16*, 312–325. [[CrossRef](#)] [[PubMed](#)]
2. Bourla, A.B.; Meropol, N.J. Bridging the Divide between Clinical Research and Clinical Care in Oncology: An Integrated Real-World Evidence Generation Platform. *Digit. Health* **2021**, *7*, 20552076211059975. [[CrossRef](#)] [[PubMed](#)]

3. Beacher, F.D.; Mujica-Parodi, L.; Gupta, S.; Ancora, L.A. Machine Learning Predicts Outcomes of Phase III Clinical Trials for Prostate Cancer. *Algorithms* **2021**, *14*, 147. [[CrossRef](#)]
4. Berger, M.L.; Curtis, M.D.; Smith, G.; Harnett, J.; Abernethy, A.P. Opportunities and Challenges in Leveraging Electronic Health Record Data in Oncology. *Future Oncol.* **2016**, *12*, 1261–1274. [[CrossRef](#)]
5. Amoroso, N.; Pomarico, D.; Fanizzi, A.; Didonna, V.; Giotta, F.; La Forgia, D.; Latorre, A.; Monaco, A.; Pantaleo, E.; Petruzzellis, N.; et al. A Roadmap Towards Breast Cancer Therapies Supported by Explainable Artificial Intelligence. *Appl. Sci.* **2021**, *11*, 4881. [[CrossRef](#)]
6. Mitsala, A.; Tsalikidis, C.; Pitiakoudis, M.; Simopoulos, C.; Tsaroucha, A.K. Artificial Intelligence in Colorectal Cancer Screening, Diagnosis and Treatment. A New Era. *Curr. Oncol.* **2021**, *28*, 1581–1607. [[CrossRef](#)]
7. Da Silva Barros, M.H.L.F.; Alves, G.O.; Souza, L.M.F.; da Silva Rocha, E.; de Oliveira, J.F.L.; Lynn, T.; Sampaio, V.; Endo, P.T. Benchmarking Machine Learning Models to Assist in the Prognosis of Tuberculosis. *Informatics* **2021**, *8*, 27. [[CrossRef](#)]
8. Kreimeyer, K.; Foster, M.; Pandey, A.; Arya, N.; Halford, G.; Jones, S.F.; Forshee, R.; Walderhaug, M.; Botsis, T. Natural Language Processing Systems for Capturing and Standardizing Unstructured Clinical Information: A Systematic Review. *J. Biomed. Inform.* **2017**, *73*, 14–29. [[CrossRef](#)]
9. Wang, Y.; Wang, L.; Rastegar-Mojarad, M.; Moon, S.; Shen, F.; Afzal, N.; Liu, S.; Zeng, Y.; Mehrabi, S.; Sohn, S.; et al. Clinical Information Extraction Applications: A Literature Review. *J. Biomed. Inform.* **2018**, *77*, 34–49. [[CrossRef](#)]
10. Yim, W.; Yetisgen, M.; Harris, W.P.; Kwan, S.W. Natural Language Processing in Oncology: A Review. *JAMA Oncol.* **2016**, *2*, 797–804. [[CrossRef](#)]
11. Savova, G.K.; Tseytlin, E.; Finan, S.; Castine, M.; Miller, T.; Medvedeva, O.; Harris, D.; Hochheiser, H.; Lin, C.; Chavan, G.; et al. DeepPhe: A Natural Language Processing System for Extracting Cancer Phenotypes from Clinical Records. *Cancer Res.* **2017**, *77*, e115–e118. [[CrossRef](#)] [[PubMed](#)]
12. Birnbaum, B.; Nussbaum, N.; Seidl-Rathkopf, K.; Agrawal, M.; Estevez, M.; Estola, E.; Haimson, J.; He, L.; Larson, P.; Richardson, P. Model-Assisted Cohort Selection with Bias Analysis for Generating Large-Scale Cohorts from the EHR for Oncology Research. *arXiv* **2020**, arXiv:2001.09765.
13. Maarseveen, T.D.; Maurits, M.P.; Niemantsverdriet, E.; van der Helm-van Mil, A.H.M.; Huizinga, T.W.J.; Knevel, R. Handwork Vs Machine: A Comparison of Rheumatoid Arthritis Patient Populations as Identified from EHR Free-Text by Diagnosis Extraction through Machine-Learning Or Traditional Criteria-Based Chart Review. *Arthritis Res. Ther.* **2021**, *23*, 174. [[CrossRef](#)]
14. Hu, Z.; Melton, G.B.; Moeller, N.D.; Arsoniadis, E.G.; Wang, Y.; Kwaan, M.R.; Jensen, E.H.; Simon, G.J. Accelerating Chart Review using Automated Methods on Electronic Health Record Data for Postoperative Complications. In *AMIA Annual Symposium Proceedings*; American Medical Informatics Association: Bethesda, MD, USA, 2016; pp. 1822–1831.
15. Collins, G.S.; Moons, K.G.M. Reporting of Artificial Intelligence Prediction Models. *Lancet* **2019**, *393*, 1577–1579. [[CrossRef](#)]
16. Sounderajah, V.; Ashrafian, H.; Golub, R.M.; Shetty, S.; De Fauw, J.; Hooft, L.; Moons, K.; Collins, G.; Moher, D.; Bossuyt, P.M.; et al. Developing a Reporting Guideline for Artificial Intelligence-Centred Diagnostic Test Accuracy Studies: The STARD-AI Protocol. *BMJ Open* **2021**, *11*, e047709. [[CrossRef](#)] [[PubMed](#)]
17. Vasey, B.; Clifton, D.A.; Collins, G.S.; Denniston, A.K.; Faes, L.; Geerts, B.F.; Liu, X.; Morgan, L.; Watkinson, P.; McCulloch, P.; et al. DECIDE-AI: New Reporting Guidelines to Bridge the Development-to-Implementation Gap in Clinical Artificial Intelligence. *Nat. Med.* **2021**, *27*, 186–187.
18. Sendak, M.P.; Gao, M.; Brajer, N.; Balu, S. Presenting Machine Learning Model Information to Clinical End Users with Model Facts Labels. *NPJ Digit. Med.* **2020**, *3*, 41. [[CrossRef](#)]
19. Hernandez-Boussard, T.; Bozkurt, S.; Ioannidis, J.P.A.; Shah, N.H. MINIMAR (MINimum Information for Medical AI Reporting): Developing Reporting Standards for Artificial Intelligence in Health Care. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 2011–2015. [[CrossRef](#)]
20. Bates, D.W.; Auerbach, A.; Schulam, P.; Wright, A.; Saria, S. Reporting and Implementing Interventions Involving Machine Learning and Artificial Intelligence. *Ann. Intern. Med.* **2020**, *172*, S137–S144. [[CrossRef](#)]
21. Girman, C.J.; Ritchey, M.E.; Lo Re, V., III. Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision-Making for Drug and Biological Products. *Pharmacoepidemiol. Drug Saf.* **2022**, *31*, 717–720. [[CrossRef](#)]
22. Mullainathan, S.; Vogeli, C.; Powers, B.; Obermeyer, Z. Dissecting Racial Bias in an Algorithm used to Manage the Health of Populations. *Science* **2019**, *366*, 447–453.
23. Willeminck, M.J.; Koszek, W.A.; Hardell, C.; Wu, J.; Fleischmann, D.; Harvey, H.; Folio, L.R.; Summers, R.M.; Rubin, D.L.; Lungren, M.P. Preparing Medical Imaging Data for Machine Learning. *Radiology* **2020**, *295*, 4–15. [[CrossRef](#)] [[PubMed](#)]
24. Varoquaux, G.; Cheplygina, V. Machine Learning for Medical Imaging: Methodological Failures and Recommendations for the Future. *NPJ Digit. Med.* **2022**, *5*, 48. [[CrossRef](#)] [[PubMed](#)]
25. Tan, W.K.; Heagerty, P.J. Surrogate-Guided Sampling Designs for Classification of Rare Outcomes from Electronic Medical Records Data. *Biostatistics* **2020**, *23*, 345–361. [[CrossRef](#)]
26. Figueroa, R.L.; Zeng-Treitler, Q.; Kandula, S.; Ngo, L.H. Predicting Sample Size Required for Classification Performance. *BMC Med. Inform. Decis. Mak.* **2012**, *12*, 8. [[CrossRef](#)]
27. Rokem, A.; Wu, Y.; Lee, A. Assessment of the Need for Separate Test Set and Number of Medical Images Necessary for Deep Learning: A Sub-Sampling Study. *bioRxiv* **2017**, 196659. [[CrossRef](#)]
28. Lakens, D. Sample Size Justification. *Collabra Psychol.* **2022**, *8*, 33267.

29. Kelly, C.J.; Karthikesalingam, A.; Suleyman, M.; Corrado, G.; King, D. Key Challenges for Delivering Clinical Impact with Artificial Intelligence. *BMC Med.* **2019**, *17*, 195.
30. Lash, T.L.; Fox, M.P.; MacLehose, R.F.; Maldonado, G.; McCandless, L.C.; Greenland, S. Good Practices for Quantitative Bias Analysis. *Int. J. Epidemiol.* **2014**, *43*, 1969–1985. [[CrossRef](#)]
31. Agrawal, M.; Adams, G.; Nussbaum, N.; Birnbaum, B. TIFTI: A Framework for Extracting Drug Intervals from Longitudinal Clinic Notes. *arXiv* **2018**, arXiv:1811.12793.
32. Jager, K.J.; van Dijk, P.C.; Zoccali, C.; Dekker, F.W. The Analysis of Survival Data: The Kaplan–Meier Method. *Kidney Int.* **2008**, *74*, 560–565. [[CrossRef](#)] [[PubMed](#)]
33. US Food and Drug Administration. *Framework for FDA'S Real-World Evidence Program*; US Department of Health and Human Services Food and Drug Administration: Silver Spring, MD, USA, 2018.
34. Desai, K.; Chandwani, S.; Ru, B.; Reynolds, M.; Christian, J.B.; Estiri, H. PCN37 an Oncology Real-World Data Assessment Framework for Outcomes Research. *Value Health* **2021**, *24*, S25. [[CrossRef](#)]
35. National Academies of Sciences, Engineering, and Medicine Division; Health and Medicine Division; Board on Health Sciences Policy. *Forum on Drug Discovery, Development, and Translation. Examining the Impact of Real-World Evidence on Medical Product Development*; Shore, C., Gee, A.W., Kahn, B., Forstag, E.H., Eds.; National Academies Press: Washington, DC, USA, 2019.
36. Hernandez-Boussard, T.; Monda, K.L.; Crespo, B.C.; Riskin, D. Real World Evidence in Cardiovascular Medicine: Ensuring Data Validity in Electronic Health Record-Based Studies. *J. Am. Med. Inform. Assoc.* **2019**, *26*, 1189–1194. [[CrossRef](#)] [[PubMed](#)]
37. Miksad, R.A.; Abernethy, A.P. Harnessing the Power of Real-World Evidence (RWE): A Checklist to Ensure Regulatory-Grade Data Quality. *Clin. Pharmacol. Ther.* **2018**, *103*, 202–205. [[CrossRef](#)] [[PubMed](#)]
38. Suresh, H.; Guttag, J.V. A Framework for Understanding Sources of Harm Throughout the Machine Learning Life Cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*; Association for Computing Machinery: New York, NY, USA, 2021.