2

Development and Experience with Cancer Risk Prediction Models Using Federated Databases and Electronic Health Records

Limor Appelbaum[1] • Irving D. Kaplan[1] • Matvey B. Palchuk[2], Steven Kundrot[2] • Jessamine P. Winer-Jones[2] • Martin Rinard[3]

[1]Department of Radiation Oncology, Beth Israel Deaconess Medical Center, Boston, MA, USA; [2]TriNetX, LLC, Cambridge, MA, USA; [3]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA

Author for correspondence: Limor Appelbaum, Department of Radiation Oncology, Beth Israel Deaconess Medical Center, Boston, MA, USA. Email: lappelb1@bidmc.harvard.edu

Cite this chapter as: Appelbaum L, Kaplan ID, Palchuk MB, Kundrot S, Winer-Jones JP, Rinard M. Development and experience with cancer risk prediction models using federated databases and electronic health records. In: Linwood SL, editor. *Digital Health*. Brisbane (AU): Exon Publications. Online first 2022 Feb 02.

Doi: https://doi.org/10.36255/exon-publications-digital-health-federated-databases

Abstract: Early diagnosis is critical to improving survival rates of lethal cancers, such as pancreatic duct adenocarcinoma (PDAC). However, there are no reliable screening test for these cancers. In this chapter, we present potential methods for predicting early, evolving cancers by leveraging readily available electronic health record (EHR) data and machine learning. We discuss the various aspects of our collaborative experience, involving clinical and computer scientists, in navigating the process of using EHRs to develop cancer risk prediction models. This chapter is intended to serve as a guide to others preforming this type of research.

In: Linwood SL, editor. *Digital Health*. Exon Publications, Brisbane, Australia.
ISBN: 978-0-6453320-1-8. Doi: https://doi.org/10.36255/exon-publications-digital-health

We cover the different steps involved, based on our initial experience of model development using single-institution data, including data acquisition, querying and downloading data, protecting patient confidentiality, data curation, model development, and validation. Challenges encountered when using single-institution data is presented, along with lessons learned. Drawing from our experience working with a federated database of EHR data from multiple institutions to develop a risk prediction model for PDAC, we also discuss how many of these challenges can be addressed by using such a federated database of EHR data. We also discuss future clinical opportunities that may arise from leveraging data from a federated network, such as the deployment of risk models for clinical studies.

Keywords: Cancer risk prediction models; electronic health records; federated network; machine learning; pancreatic duct adenocarcinoma

INTRODUCTION

Early diagnosis of cancer is critical to delivering the best outcomes for patients. General population screening leading to early diagnosis is well established for breast and colon cancer (1) but lacking for other, relatively rare but lethal cancers, such as pancreatic, ovarian, and gastric adenocarcinomas. For the minority of high-risk individuals eligible for screening, e.g., individuals with an inherited predisposition to PDAC, it has been shown to be crucial to improving survival rates (2). The absence of general population screening for these malignancies is mostly due to the high cost of screening the general public for low prevalence cancers, using expensive modalities such as imaging, compounded by the lack of highly-effective strategies that meet the expected performance metrics for a screening tool (3). Clinical and laboratory data are routinely collected in EHRs as part of patient clinical encounters. Real-world data, such as that found in EHRs, is observational data, i.e., not collected specifically for a research study, as opposed to data gathered in an experimental setting, such as a randomized-controlled trial (RCT) (4). This data is widely available, contains a recording of the patient's medical history, and is rich in information on their clinical state over time (5). It has been shown that various signs, symptoms, and risk factors can precede cancer diagnosis, e.g., weight loss and diabetes in PDAC (6, 7). These are reflected in data elements found in EHRs such as diagnoses, medications, and labs and can be leveraged to identify individuals at increased risk of cancer development in the future.

In this chapter, we discuss the development of cancer prediction models as a possible solution for expanding the indications for screening, whether as standalone tools in the clinic, or as part of a tiered system leading to early diagnosis. We describe the rationale for using EHR data for developing cancer risk prediction models in terms of its widespread availability and how development in the same setting can facilitate clinical implementation. Next, we outline our approach to leveraging EHR data to develop cancer risk prediction models. We also explain why we chose a data-driven approach over an approach that predefines specific features for the model, and how using machine learning techniques to discover less obvious features can improve model performance over using 'conventional' feature sets. We then summarize the necessary steps needed to be taken to acquire and use the data, the challenges we encountered during these steps, and what can be done to minimize these steps and reduce the time to model deployment (federated network). Throughout the chapter, we give examples from our work on developing and validating models for the prediction of PDAC.

THE GOAL OF DEVELOPING CANCER RISK PREDICTION MODELS

Risk prediction models are simple, inexpensive, and non-invasive tools for identifying individuals at higher risk for a specific cancer than the general population. They can serve to close the gap between the need for more screening and early diagnosis in relatively rare but highly lethal cancers, and the current high costs of screening, by identifying high-risk cohorts out of the general population. For rare types of cancer, in which population-wide screening is not cost-effective, risk prediction models can be used to identify high-risk sub-populations that would benefit most from surveillance and early diagnosis. Depending on their performance metrics and the cancer-specific screening modality, risk prediction models can be used as a standalone tool for determining which individuals are high-risk and need screening for a certain cancer or as part of a tiered system (e.g., together with biomarker testing) (8).

Over the past decade, numerous cancer risk prediction models have been published. However, few have progressed past the development and initial validation stages to prospective real-time validation in the USA. The reasons behind this are complex, but much can be learned from the attempts at implementation of similar models in Europe, where several risk assessment tools have been integrated into practice, most notably the 'QCancer' and Risk Assessment Tool (RAT) models (9). These were tested as decision support tools based solely on a generated risk score, for multiple cancers (including PDAC and ovarian adenocarcinoma), and while reporting fair performance, their adoption by primary care physicians has been slow (9). These models are likely limited by their "provider dependency," or varied general practitioner interpretation of patient symptoms leading to inconsistent estimates by the model, as well as by design and integration of the tool (10). There are also no reports of these models continuously generating an updated risk score based on new data as it becomes available (9).

Our approach, outlined in the following sections, is to develop models with a specific use-case scenario in mind. Therefore, we have attempted to develop models that are interpretable, generalizable, and easily reproducible. We also believe making them simple to integrate and automated, continuously incorporating new data as it is entered into the system, will ease their clinical implementation. A real-time prospective validation is a key to assessing these factors and determining the model's true performance.

THE RATIONALE FOR USING REAL-WORLD DATA TO BUILD MODELS

Cancer risk prediction models can be developed using almost any type of clinical data. Numerous published models have been developed using data from RCTs, survey data, or more commonly, population cohort studies (11). These data sources have the advantage of having 'full' data for the patients included, however they represent highly selected populations and are therefore more likely to be non-generalizable. In contrast, developing these models on real-world data, or data not collected for a study, but routinely collected as part of daily practice, has several advantages. First, the patient medical record captures the entire 'patient journey,' so it contains changes in the patient's clinical state over time as recorded by their providers. These changes are reflected in diagnoses, lab tests, medications, etc. For example, many individuals that will develop PDAC in the future may experience weight loss, changes in blood glucose levels, and medication changes (initiation of glucose-lowering drugs), which are captured in their diagnostic codes, blood test results, and medication list. In addition, EHR data is available to some extent for every person using the healthcare system and not a select subgroup, as is the case when using data from an RCT. This approach for data collection is likely to be more representative of the real-world diverse patient population.

Since this data is routinely collected when an individual interacts with the healthcare system, it is readily available and exists in every healthcare system that uses EHRs. Therefore, developing cancer prediction models using this data does not involve collecting new data but rather capitalizing on the massive amounts of information routinely collected at the point of care. Lastly, building cancer prediction models leveraging EHR data is logical when envisioning their future use-case scenario. Developing the model in the same setting in which it will be clinically implemented, i.e., in the physician's office *within* the electronic patient medical record, deals upfront with some of the challenges inherent to this type of data source and helps overcome these challenges during the development and testing stages (12, 13).

MAXIMIZING THE POTENTIAL OF EHR DATA FOR MODEL DEVELOPMENT

EHR data has several well-known inherent challenges. For example, since this data is not collected at a pre-planned time or by pre-planned methods, it is non-uniform and has missing data points. However, these inherent qualities, and its large variance in types of data collected, can also offer advantages when developing cancer prediction models (14).

Feature selection

Different approaches to feature selection can be taken. In general, features can be predefined based on domain experts and/or literature review (15). Alternatively, a data-driven approach can be utilized, in which all available features are initially included, and feature selection occurs during the model development phase (16). In developing our PDAC risk prediction models, we selected machine learning models and a data-driven approach to feature selection for our models. For both our preliminary model, which was based on EHR data from two local hospitals, and our later models, based on federated network data, we experimented with a variety of model classes, including logistic regression models, neural networks, random forest, and XGBoost. Our goal was to test if these algorithms could find 'patterns' in the data of individuals that would develop PDAC in the future.

Instead of initially predefining specific features for the model to use in its predictions, as has often been done by others (17), we incorporated all routine data features that were available in our EHR datasets, and let the model 'decide' which features it needed to maximize its predictive performance. For our logistic regression models, we relied on regularization to address the potential risk of overfitting. All routine data refers to all demographic, diagnoses, medications, and lab test data that were available for each patient were included in the analysis. When examining the features our logistic regression models leveraged for their PDAC predictions, we found that previously known risk factors, signs, and symptoms, such as smoking status, weight loss, and diabetes, were given high absolute weights (18). Hence these known features were important drivers of overall model performance, lending credence to the validity of the models. However, there were many novel features with predictive value, such as hypertension, hyperlipidemia, and heart disease, that we would not have chosen to include a priory had we preselected the features. Indeed, the logistic regression models we developed leveraged thousands of different features to reach maximal performance in predicting PDAC. Experiments conducted in an attempt to reduce the number of features included in the model resulted in decreased model performance.

Exploring 'unconventional' features

EHR data represent not only the clinical state of the patient but, importantly, are also a reflection of the interaction of the patient with the healthcare system (14). Therefore, the timing and frequency of events within the patient medical record could be clinically significant and potentially influence model performance (14). For example, we conducted experiments exploring different characteristics of the various available feature sets (diagnoses, labs, medications, demographics) and found that the frequency of administration of certain lab tests per patient was a valuable feature for improving the performance of the model. For almost every type of lab test, the average number of administrations per patient was higher for the patients who later developed PDAC compared to the controls. The lab tests with the highest discrimination included glucose, hemoglobin, hematocrit, and creatinine.

Handling missing data

Missing data is an issue that one must deal with when building models using EHR data. The first question that arises is what accounts for 'missing' data? Is the data missing completely at random, or is that missingness clinically meaningful, and can it be capitalized upon in developing the model? If the data is missing-at-random (or one assumes that assume it is), various methods of imputation, including single imputation, multiple imputation, multivariate imputation by chained equations (MICE), and others (19, 20), can be used to fill in the missing values, based for example, on the average for those cases with an existing variable. In the previous example, we found that the number of specific lab tests administered (creatinine, hemoglobin, etc.) was much higher for those that developed PDAC versus those that did not. This can also be viewed as a type of 'missing' data but one that actually helps prediction.

Our approach to missing data was to first attempt to determine whether the missing data was at random or not. Then, based on that determination, we used imputation to fill in the gaps for some of our missing data and capitalized on that 'missingness' for others. In contrast, completely removing those cases with missing values could potentially introduce bias into risk prediction models. Removal of these cases can be viewed similarly to an analysis of only data from "perfect" subjects (Per Protocol Analysis) versus the gold standard Intent-to-Treat Analysis, where the analysis is based on all subjects that were initially enrolled in an RCT (21). Removing cases due to missing data introduces significant issues of health care inequalities.

Model interpretability

Another important consideration when using EHR data to develop cancer prediction models is to make the components of the final model interpretable to the user, in this case, the patient's provider. Since much of the routine data collected in EHRs is structured data, as is the case for demographic data, diagnostic codes, medications, etc., it is relatively simple to use logistic regression to decipher the different data elements or features that are used by the model. Knowledge of the features that 'drive' the model has been shown to increase physician trust and use of these models (22), which is critical to their future clinical implementation.

OUR INITIAL EXPERIENCE WITH LOCAL HOSPITAL DATA

Several crucial steps must be accomplished prior to analyzing the EHR data to develop a cancer risk prediction model.

Institutional review board (IRB) approval

The first of these steps is obtaining IRB approval to utilize the data for model development. This data is typically retrospective and can go through the exempt review process, since it involves use of existing data that will be de-identified. However, if the source of the data is another institution, or the data analysis is being done by outside collaborators, approval to use the data may be less straightforward. It is important to note that in these cases, a Data Use Agreement (DUA) may need to be in place before this data can be accessed, and approval by a Data Sharing Committee may also be required. In our case, we initially received permission via the IRB exempt pathway to use our institutional data to develop the model. However, to externally validate our model on a different dataset from another health care system, we were required to go through the process of obtaining not only IRB approval for using that hospital's data, but also the approval of the Data Sharing Committee, and a DUA. We consider validation on new data that was not used for model training (external validation) to be a crucial step in building reproducible and generalizable models. This process took over one year to complete and is, in our experience, one of the biggest hurdles in working with multi-institution data.

Downloading the data, case selection and validation

The next step before data analysis is querying and downloading the data from the hospital EHR database. Many institutions have a query tool in place though which the cancer cohort, as well as the control group, can be selected. To develop our PDAC prediction model, we initially selected our cases using ICD (International Classification Codes developed by the World Health Organization) codes for PDAC. We manually validated about 100 cases to confirm that they were true cases of PDAC before model development. To our surprise, only about 40% of the ICD coded PDAC cases were real cases, others were different cancer types, and some were pancreatic disease, but not cancer. Therefore, to ensure model training on confirmed PDAC cases, ICD coded PDAC patients were cross-checked against the hospital tumor registry: patients with no corresponding registry record were removed. Based on this method, about 40% of ICD-coded PDAC cases were accurate. Since data in these registries are audited and must meet CDC standards (https://www.cdc.gov/cancer/npcr/index.htm), this is the 'gold-standard' for validation of a cancer diagnosis. We emphasize that this confirmation is only needed when developing the model and would not be required in its eventual use-case scenario since the model would then be assessing the data of healthy individuals. To clarify, data from PDAC patients is needed for the model to 'learn' the patterns that will occur in people that will develop the disease. However, when used in the clinic, it will be deployed on the data of healthy individuals and will identify those patterns to predict future PDAC. We then selected a control group by excluding all patients with a history of PDAC and matched them to cases by age and sex.

Protecting patient confidentiality

The next step is ensuring patient confidentiality by de-identification of all patient data. De-identification can be done using one of two methods in accordance with the Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule: the Safe-Harbor method or Expert Determination. For the development dataset, we used the Safe-Harbor method to de-identify our data. In addition, to avoid the need for our computer scientist collaborators to store the data on their own computing infrastructure, we investigated various mechanisms that would allow the data to remain behind the hospital firewall but still have our collaborators access it to develop the model. Our solution was to establish a virtual machine (VM), set up specifically for our project by Academic Computing/Information Systems. The VM allowed the computer scientists access to the de-identified data, without the data ever ‘leaving’ the hospital. We first de-identified the data, removing all Protected Health Information (PHI), and stored it on a network share file specific to the project, behind the hospital firewall, from which it was accessible via the VM for analysis. It was also necessary for our collaborators to acquire hospital collaborator status before being given access to hospital data.

Data preparation

The next step was to transform the raw data into useable data. This ‘cleaning’ or data preparation process includes standardizing and normalizing the data. Example cleaning activities included standardizing drug names present as generic and various trade names into one common format and removing extreme outliers of lab test values, i.e., those outside the normal and reasonable pathologic values for disease. The latter can be done using various previously described methods (23). The criteria we used for defining outliers was more than 3 standard deviations away from the mean for each lab test value.

This step also includes assessing the data for any inconsistencies influencing prediction. In our example, we found that many data elements were missing for cases compared to the controls (Individuals without a diagnosis of PDAC, matched for age and sex to cases) While the algorithm could use this ‘missingness’ to classify patients into a certain group, when manually inspecting some of these records, we found that this ‘missingness’ was a result of cancer cases being referred to our hospital for either a workup for suspected PDAC or immediately after cancer diagnosis for treatment. Therefore, they did not have a history at our hospital, and lacked the detailed EHR data needed for the model. We addressed this discrepancy by adding a filter that would remove patients without a ‘history’ at our hospital, ensuring that included patients would have enough data for developing the model. This example of data cleaning that we needed to do before model development, underscores the need to understand the specific data before analysis.

Model development

Following the steps outlined above, we developed our initial model using local hospital data. We experimented with various data elements, data combinations, and model classes, namely logistic regression and neural networks. We found the logistic regression models to perform comparably to neural networks, but logistic regression models had the additional advantage of ease of interpretability compared with the more complex neural networks. For our initial model, we found diagnostic codes to be very good on their own at predicting future PDAC. We next sought to externally validate our preliminary model on the new dataset, which we had acquired from another local hospital system. This involved the steps outlined above before actual model validation, including tumor registry validation of cases, de-identification on the data, and transferring it to our hospital network file sharing folder so that it could reside behind the firewall. Following these steps, the data was accessed by our collaborators for data ‘cleaning’ and preparation. We then validated our model on this data from an external source.

OVERCOMING THE LIMITATIONS OF SINGLE-INSTITUTION DATA USING A FEDERATED NETWORK DATABASE

To simplify and expedite data collection and processing as described above, we sought to streamline this process by collaborating with a global federated network company. In this section, we outline what a federated network is and the different ways in which the necessary steps preceding model building outlined above are achieved within this network.

What is a federated network?

Let us consider the EHR data from a single institution to be the atomic element. That data is informative of the state of care at that institution and the healthcare burden of the local population. Findings based on analyses of data from a single institution may be influenced by things such as institutional policies, available resources, and the characteristics of the population it serves. In addition, depending on the institution's size and prevalence of the condition of interest, it may be challenging to find sufficient patients to power the analysis.

There is an increasing desire to analyze data from multiple institutions. There are many advantages to this capability, for example larger and more data sets and an increased ability to develop and validate models that work across multiple institutions. The challenge lies in finding a secure and compliant, but useful and sustainable, way to share data across institutions. The two main approaches are aggregation and federation.

In an aggregated network, data is physically brought together in a single location (database, data lake, etc.). With this approach, the data becomes easier to analyze and useful for tasks that require large amounts of processing power, such as training machine learning models. The challenge is that many institutions are reluctant to give up control of their data in this way. In particular, it has become progressively more difficult to use an aggregation approach in ex-US countries as more and more enact legislative privacy protection frameworks based on or similar to GDPR.

One of the better-known aggregated repositories of clinical patient data created in response to COVID-19 pandemic is the National COVID Cohort Collaborative (N3C) (24). Created under the auspices of the National Center for Advancing Translational Sciences (NCATS) at the National Institutes of Health (NIH), N3C is a platform for collecting and analyzing data on COVID-19 patients and controls. The data is aggregated in a secure enclave, where it is harmonized and made available to researchers. As of this writing, 69 sites have contributed data to this effort, and the repository contains ~11.1 billion rows. N3C focuses exclusively on COVID-19, and the goal of the project was to respond to this emergency as quickly as possible. This may explain the success of the aggregated approach in this particular use case.

In a federated network, the data remains at the contributing institution or in a cloud-based instance under their control and what travels across such network are questions or queries going outbound to the data and answers are coming back. In essence, the question is brought to the data rather than the data being brought to the question. To the user, the data appears as a single database, but this database is virtual. In fact, there are multiple databases operating in concert behind the scenes.

For a federated network to function, the data needs to be harmonized both syntactically (structure) and semantically (meaning), and a common "language" must be agreed upon for queries and responses. In other words, participating systems must be interoperable—be able to meaningfully share the data and be able to act upon it.

TriNetX (25), a commercial company founded in 2013 and based in Cambridge, MA, has built and is operating the largest global research network of real-world clinical data. The TriNetX platform consists of a federated data repository and a growing set of rich cohort identification and data analytics capabilities that allow real-time data access to academic and private industry researchers. A private-public partnership forms the core of this network as the majority of clinical data is contributed by healthcare organizations, many of which are large academic medical centers, while the funding comes from subscriptions offered to Life Science companies for access to the data, analytics, and services.

What kind of approvals are required to use this data?

Regardless of the form the network takes, aggregated, or federated, data sharing must be done in a way that preserves patient privacy, maintains data security, and complies with all legal regulations and contractual obligations. The specific requirements for and restrictions on data sharing vary by institution, nation, and region; however, in general, data can only be shared as a de-identified, pseudo-anonymized, or limited data set. Part of the process of combining data into a multi-institutional repository such as TriNetX or N3C is obtaining the necessary IRB approvals or waivers to allow for further sharing and use of the data under defined conditions. This alleviates the burden on individual researchers to obtain IRB approvals or waivers from each contributing institution.

How does the data querying and downloading process work?

To increase the versatility of the data, healthcare organizations contributing to TriNetX are grouped into Networks based on characteristics such as geography and data use permissions. Investigators with the appropriate institutional access rights can interrogate the federated Networks via a web-based portal. Queries are constructed in the web application and passed down to the individual physical or cloud-based data repositories. The institutional results are then passed back to the web application and aggregated into a final summary count. Depending on the specific network, investigators may have access to advanced analytic tools built into the web application or the ability to request a data download for use with other analytic software applications. When requesting a data download, investigators construct one or more queries that identify the population(s) of interest. This request is then submitted to TriNetX, where it is evaluated to ensure it complies with all applicable laws, policies, and contracts, and a data use contract specific to the project is prepared. Once the contract is finalized, the dataset is prepared by TriNetX and transferred securely to the investigator.

How is patient confidentiality maintained?

Data in TriNetX platform is functionally de-identified based on expert attestation. HIPAA allows two main ways of ensuring de-identification—via safe harbor (removing proscribed PHI elements) or via expert attestation. The expert performs formal statistical analysis to ensure the risk of re-identification has been sufficiently minimized. Any recommendations by the expert aimed at safeguarding patient privacy are implemented. For example, certain concepts which have a high likelihood of appearing in public data sources such as newspapers (diagnoses related to car accidents, misadventures suffered by astronauts, etc.) are eliminated from the interface terminology of the platform and, therefore, are not accessible.

How is the data standardized/normalized?

The methods of documenting any specific clinical fact vary over time, by healthcare organization, and by regional and national regulations. Two major challenges must be overcome to turn disparate data sources into a harmonized federated network that a novice user can easily interrogate. First, the data must be ingested into a common data model (CDM) to ensure syntactic standardization. Second, each method of documenting a clinical concept needs to be mapped to a standardized interface terminology to achieve semantic standardization.

To overcome the first challenge, the desired data is identified within the EHR, existing data warehouses, or research repositories. Using a collection of custom connectors and toolkits, this data can then be ingested into the TriNetX CDM from a range of other CDMs and healthcare information exchange standards. Currently, TriNetX can ingest data from a variety of sources, including i2b2, OMOP, HL7, and a selection of EHR systems.

For the second challenge, TriNetX maps individual site data to the TriNetX interface terminology. The mappings are constructed using a combination of site-specific custom mappings, existing public and private code set mappings (https://www.nlm.nih.gov/research/umls/mapping_projects/snomedct_to_icd10cm.html), (https://loinc.org/relma/), custom code set (26, 27) mappings, and string matching . In some cases, additional data manipulation may be required, such as unit conversion for lab results and vital signs, and mapping local text results (e.g., “detected” and “not detected”) to a set of standard results (e.g., “positive” and “negative”). This process is made significantly more complex as new countries are added, many with their own code set variations, approved medications, language, and approach to documentation.

The main strength of using a single interface terminology is it reduces the burden on the investigator and lowers the barrier to performing analysis on a global dataset. For example, when interacting with the platform, a researcher only needs to know the RxNorm code for the medication of interest in order to identify patients across the network with that medication.

However, there are several challenges and limitations to this approach. First, mappings are never perfect, and a degree of specificity and completeness may be lost in the process. Second, certain concept domains, such as procedures, are particularly challenging because they lack a globally adopted standard (28), and custom mappings are required for each new country added to the network (26). Third, the process of building custom concept maps is labor-intensive and requires a periodic refresh as new terms are added. As a result, it is not possible to map every term to the interface terminology, and there may be a delay between the introduction of a new term and when it becomes available. However, when there is a pressing clinical need, such as the onset of the COVID-19 pandemic, new terms can be quickly added and mapped across all appropriate terminologies.

Steps involved in model development using TriNetX data

As evident from the above, using data from a federated network, such as TriNetX, expedites the pre-analysis phase. Members of TriNetX network comprise two distinct groups - researchers who represent healthcare organizations and researchers who represent life science companies such as pharma or contract research organizations (CRO). Healthcare organizations share their data and therefore there is no additional cost for their participation on the network (one exception is when data from outside of their home institution is being purchased for a grant-funded study) and utilization of the platform's capabilities. Life science companies purchase subscriptions to get access.

One agreement is required to access data from multiple institutions, patient data is de-identified, and much of the data is already cleaned and prepared for analysis (e.g., standardization of ICD and medications, normalization of lab values, etc.).

Following the execution of the collaboration agreement, we used the query tool to select cases and controls and received the data as .csv files, containing all the data elements for each individual in the query (all diagnostic codes, medications, labs, etc). Our next steps included case validation using tumor registry/ pathology data, case-control matching, and handling missing data, followed by model training and testing. Finally, we focused on extending our initial model using the TriNetX database by trying various feature combinations and exploring different machine learning model classes.

OPPORTUNITIES FOR REAL-TIME MODEL DEPLOYMENT AND ASSESSMENT

In this section we discuss clinical opportunities that may arise from leveraging data from a federated network.

Build your own model (BYOM)- what is it, how does it work?

TriNetX has developed a set of functionalities that will allow an arbitrary model to be executed on a selected scope of data (select Health Care Organizations). The results of this model – typically a calculated value – will then be available for use in defining a patient cohort of interest. For example, although ICD-10 greatly expanded upon ICD-9, many rare diseases do not have a specific diagnosis code, and identifying these patients via standard query methods can be challenging. With BYOM, researchers can train a supervised machine learning model using a downloaded dataset of other data in the patient's record to calculate the likelihood of this patient having the diagnosis of interest. This likelihood score can then be deployed on the platform and used in query development as part of the inclusion or exclusion criteria. Models could range from simple algebraic calculations based on visit history, to linear regression models, to deep neural nets.

Using BYOM for a prospective clinical trial

After training and testing our PDAC prediction models, we sought to integrate our best-performing model into the federated network platform. We used information available on the platform to search for partner institutions with both primary care providers within their institution and large numbers of treated PDAC patients. We then reached out to and formed collaborations with three institutions that are part of the network and were interested in participating in a prospective study evaluating the model.

An internal network was built within the platform for the data of the three collaborating institutions. A script was developed, which can be integrated into the network, and through which the machine learning model will be deployed. This script generates a risk score for each patient, which is saved in the platform. Patient risk scores linked to their de-identified codes can then be used to prospectively validate the model in real-time, assessing model discriminatory performance and calibration. These scores can also be accessed by the PIs at collaborating institutions, allowing them to re-identify subjects assigned by the model to the high-risk group and enroll interested individuals for blood collection and biomarker analysis.

CONCLUSION

Federated networks can facilitate the development of generalizable cancer risk prediction models using multi-institutional data and streamline their integration into the EHR systems of multiple institutions. This integration allows prospective model validation, the establishment of clinical studies acting on model output, and amplifies our ability to identify at-risk cohorts and expand the indications for screening for lethal cancers.

Acknowledgement: This research was supported by funding from the Prevent Cancer Foundation.

Conflict of Interest: Matvey Palchuk, Steve Kundrot, and Jessamine Winer-Jones are employees of TriNetX. The other authors declare no potential conflicts of interest with respect to research, authorship and/or publication of this manuscript.

Copyright and Permission Statement: The authors confirm that the materials included in this chapter do not violate copyright laws. Where relevant, appropriate permissions have been obtained from the original copyright holder(s), and all original sources have been appropriately acknowledged or referenced.

REFERENCES

1. Siu AL, Force USPST. Screening for Breast Cancer: U.S. Preventive Services Task Force Recommendation Statement. Ann Intern Med. 2016;164(4):279–96. https://doi.org/10.7326/M15-2886
2. Canto MI, Almario JA, Schulick RD, Yeo CJ, Klein A, Blackford A, et al. Risk of Neoplastic Progression in Individuals at High Risk for Pancreatic Cancer Undergoing Long-term Surveillance. Gastroenterology. 2018;155(3):740–51 e2. https://doi.org/10.1053/j.gastro.2018.05.035
3. Kenner B, Chari ST, Kelsen D, Klimstra DS, Pandol SJ, Rosenthal M, et al. Artificial Intelligence and Early Detection of Pancreatic Cancer: 2020 Summative Review. Pancreas. 2021;50(3):251–79. https://doi.org/10.1097/MPA.0000000000001762
4. Khozin S, Blumenthal GM, Pazdur R. Real-world Data for Clinical Evidence Generation in Oncology. J Natl Cancer Inst. 2017;109(11). https://doi.org/10.1093/jnci/djx187
5. Rayner J, Khan T, Chan C, Wu C. Illustrating the patient journey through the care continuum: Leveraging structured primary care electronic medical record (EMR) data in Ontario, Canada using chronic obstructive pulmonary disease as a case study. Int J Med Inform. 2020;140:104159. https://doi.org/10.1016/j.ijmedinf.2020.104159
6. Hart PA, Kamada P, Rabe KG, Srinivasan S, Basu A, Aggarwal G, et al. Weight loss precedes cancer-specific symptoms in pancreatic cancer-associated diabetes mellitus. Pancreas. 2011;40(5):768–72. https://doi.org/10.1097/MPA.0b013e318220816a
7. Pannala R, Basu A, Petersen GM, Chari ST. New-onset diabetes: a potential clue to the early diagnosis of pancreatic cancer. Lancet Oncol. 2009;10(1):88–95. https://doi.org/10.1016/S1470-2045(08)70337-1
8. Putcha G, Gutierrez A, Skates S. Multicancer Screening: One Size Does Not Fit All. JCO Precis Oncol. 2021;5:574–6. https://doi.org/10.1200/PO.20.00488
9. Medina-Lara A, Grigore B, Lewis R, Peters J, Price S, Landa P, et al. Cancer diagnostic tools to aid decision-making in primary care: mixed-methods systematic reviews and cost-effectiveness analysis. Health Technol Assess. 2020;24(66):1–332. https://doi.org/10.3310/hta24660
10. Chiang PP, Glance D, Walker J, Walter FM, Emery JD. Implementing a QCancer risk tool into general practice consultations: an exploratory study using simulated consultations with Australian general practitioners. Br J Cancer. 2015;112 Suppl 1:S77–83. https://doi.org/10.1038/bjc.2015.46
11. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc. 2017;24(1):198–208. https://doi.org/10.1093/jamia/ocw042
12. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, Bernstam EV, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care. 2013;51(8 Suppl 3):S30–7. https://doi.org/10.1097/MLR.0b013e31829b1dbd
13. Elmore JG, Barnhill RL, Elder DE, Longton GM, Pepe MS, Reisch LM, et al. Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study. BMJ. 2017;357:j2813. https://doi.org/10.1136/bmj.j2813
14. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. BMJ. 2018;361:k1479. https://doi.org/10.1136/bmj.k1479
15. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98(9):691–8. https://doi.org/10.1136/heartjnl-2011-301247
16. Richter AN, Khoshgoftaar TM. A review of statistical and machine learning methods for modeling cancer risk using structured clinical data. Artif Intell Med. 2018;90:1–14. https://doi.org/10.1016/j.artmed.2018.06.002
17. Boursi B, Finkelman B, Giantonio BJ, Haynes K, Rustgi AK, Rhim AD, et al. A clinical prediction model to assess risk for pancreatic cancer among patients with prediabetes. Eur J Gastroenterol Hepatol. 2022;34(1):33–8. https://doi.org/10.1097/MEG.0000000000002052

18. Appelbaum L, Cambronero JP, Stevens JP, Horng S, Pollick K, Silva G, et al. Development and validation of a pancreatic cancer risk model for the general population using electronic health records: An observational study. Eur J Cancer. 2021;143:19–30. https://doi.org/10.1016/j.ejca.2020.10.019
19. Waljee AK, Mukherjee A, Singal AG, Zhang Y, Warren J, Balis U, et al. Comparison of imputation methods for missing laboratory data in medicine. BMJ Open. 2013;3(8). https://doi.org/10.1136/bmjopen-2013-002847
20. Muhlenbruch K, Kuxhaus O, di Giuseppe R, Boeing H, Weikert C, Schulze MB. Multiple imputation was a valid approach to estimate absolute risk from a prediction model based on case-cohort data. J Clin Epidemiol. 2017;84:130–41. https://doi.org/10.1016/j.jclinepi.2016.12.019
21. Brody T. Clinical trials: study design, endpoints and biomarkers, drug safety, and FDA and ICH guidelines: Academic press; 2016. https://doi.org/10.1016/B978-0-12-804217-5.00025-4
22. Diprose WK, Buist N, Hua N, Thurier Q, Shand G, Robinson R. Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. J Am Med Inform Assoc. 2020;27(4):592–600. https://doi.org/10.1093/jamia/ocz229
23. Estiri H, Klann JG, Murphy SN. A clustering approach for detecting implausible observation values in electronic health records data. BMC Med Inform Decis Mak. 2019;19(1):142. https://doi.org/10.1186/s12911-019-0852-6
24. Haendel MA, Chute CG, Bennett TD, Eichmann DA, Guinney J, Kibbe WA, et al. The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. J Am Med Inform Assoc. 2021;28(3):427–43. https://doi.org/10.1093/jamia/ocaa196
25. Topaloglu U, Palchuk MB. Using a Federated Network of Real-World Data to Optimize Clinical Trials Operations. JCO Clin Cancer Inform. 2018;2:1–10. https://doi.org/10.1200/CCI.17.00067
26. Schulz S, Steffel J, Polster P, Palchuk M, Daumke P, editors. Aligning an Administrative Procedure Coding System with SNOMED CT. JOWO; 2019.
27. Millan-Fernandez-Montes A, Perez-Rey D, Hernandez-Ibarburu G, Palchuk MB, Mueller C, Claerhout B. Mapping clinical procedures to the ICD-10-PCS: The German operation and procedure classification system use case. Journal of Biomedical Informatics. 2020;109:103519. https://doi.org/10.1016/j.jbi.2020.103519
28. Maier C, Lang L, Storf H, Vormstein P, Bieber R, Bernarding J, et al. Towards Implementation of OMOP in a German University Hospital Consortium. Appl Clin Inform. 2018;9(1):54–61. https://doi.org/10.1055/s-0037-1617452