Automatic information extraction from childhood cancer pathology reports

Hong-Jun Yoon¹, Alina Peluso¹, Eric B. Durbin², Xiao-Cheng Wu³, Antoinette Stroup⁴, Jennifer Doherty⁵, Stephen Schwartz⁶, Charles Wiggins⁷, Linda Coyle⁸, and Lynne Penberthy⁹

¹Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA, ²College of Medicine, University of Kentucky, Lexington, Kentucky, USA, ³School of Public Health, Louisiana State University Health Sciences Center, New Orleans, Louisiana, USA, ⁴Rutgers Cancer Institute of New Jersey, New Brunswick, New Jersey, USA, ⁵Huntsman Cancer Institute, University of Utah, Salt Lake City, Utah, USA, ⁶Fred Hutchinson Cancer Research Center, Epidemiology Program, Seattle, Washington, USA, ⁷University of New Mexico, Albuquerque, New Mexico, USA, ⁸Information Management Services Inc., Calverton, Maryland, USA, and ⁹National Cancer Institute, Bethesda, Maryland, USA

Corresponding Author: Hong-Jun Yoon, Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, TN 37830, USA; yoonh@ornl.gov

Received 18 January 2022; Revised 9 May 2022; Editorial Decision 17 May 2022; Accepted 3 June 2022

ABSTRACT

Objectives: The International Classification of Childhood Cancer (ICCC) facilitates the effective classification of a heterogeneous group of cancers in the important pediatric population. However, there has been no development of machine learning models for the ICCC classification. We developed deep learning-based information extraction models from cancer pathology reports based on the ICD-O-3 coding standard. In this article, we describe extending the models to perform ICCC classification.

Materials and Methods: We developed 2 models, ICD-O-3 classification and ICCC recoding (Model 1) and direct ICCC classification (Model 2), and 4 scenarios subject to the training sample size. We evaluated these models with a corpus consisting of 29,206 reports with age at diagnosis between 0 and 19 from 6 state cancer registries.

Results: Our findings suggest that the direct ICCC classification (Model 2) is substantially better than reusing the ICD-O-3 classification model (Model 1). Applying the uncertainty quantification mechanism to assess the confidence of the algorithm in assigning a code demonstrated that the model achieved a micro-F1 score of 0.987 while abstaining (not sufficiently confident to assign a code) on only 14.8% of ambiguous pathology reports.

Conclusions: Our experimental results suggest that the machine learning-based automatic information extraction from childhood cancer pathology reports in the ICCC is a reliable means of supplementing human annotators at state cancer registries by reading and abstracting the majority of the childhood cancer pathology reports accurately and reliably.

Key words: pediatric cancer, cancer pathology reports, information extraction, machine learning
LAY SUMMARY
ICCC is the coding standard designed to categorize childhood cancers. However, machine learning-based ICCC classification has not been extensively studied, mainly owing to the limited volume of the pediatric cancer corpus; pediatric cancer is much less prevalent than adult cancers. Under the oversight of the National Childhood Cancer Registry project, we developed a deep learning-based text comprehension model for classifying ICCC from childhood cancer pathology reports. We performed a comparison study between (1) classifying ICD-O-3 codes and then recoding into ICCC and (2) classifying ICCC codes directly. We observed that the second approach exhibited a substantially higher accuracy score.

We are aware that the low-precision models are not appropriate for this exercise because they will degrade the credibility of the model-based decisions. We applied an uncertainty quantification algorithm to the ICCC classification model. We achieved nearly perfect accuracy scores, while the model passed over 14.8% of ambiguous cases. This result means our machine learning model can serve human annotators at state cancer registries by processing 85.2% of the childhood cancer pathology reports automatically.

INTRODUCTION
Cancer is the leading cause of death by disease in American children ages 0–19 years. Each year, nearly 16 000 children in the United States and over 300 000 children globally are diagnosed with cancer. Analysis of the population-level data for childhood cancers will increase the understanding of the factors that cause cancer and shed light on factors that may help prevent against cancer, thus providing evidence to guide public health recommendations and identify and develop improved treatments. Further rapid identification and classification of cases might be used to enhance enrollment for clinical trials for ultra-rare pediatric cancers, thereby enabling access to state-of-the-art care to a wider set of childhood cancer patients.

Cancer pathology reports are an excellent resource for such studies. A pathology report is a medical document written by a pathologist that contains the diagnosis determined by examining cells and tissues under a microscope. The reports include information about the topography (site of origin) and the morphology (histology and behavior) of the tumor. Automatic information extraction is the machine learning (ML) based method that abstracts the findings using the standardized codes.

We have been researching automatic information extraction and abstraction from cancer pathology reports based on the International Classification of Diseases for Oncology, third edition (ICD-O-3) coding standard. Recent advances in AI have enabled us to establish robust natural language processing and text comprehension algorithms, which could help mitigate the overhead of manually curating data. We have demonstrated that the deep learning (DL) models exhibited state-of-the-art performance compared against traditional ML-based and rule-based approaches. However, to the best of our knowledge, the efficacy of using the existing automatic information extraction models for childhood cancer pathology reports has not been studied.

The difficulties of applying the existing automatic information extraction model to the cancer pathology reports from childhood cancers originated from the differences of prevalent cancer types between adult and pediatric cancer cases. The most prevalent adult cancers are breast, lung and bronchus, prostate, and colorectal cancers. In contrast, the most common childhood cancers are leukemia, lymphoma, and tumors of the central nervous system. Cancers more prevalent in adults are underrepresented among the children, and—likewise—the cancers that are more prevalent in children are underrepresented among the adults. Because the number of adult cancers is substantially higher than the number of childhood cancers, the ML model trained using the entire cancer pathology data corpus is likely biased toward adult cancers. Consequently, this discrepancy may manifest downstream as a classification performance decrease for the childhood cancer model.

Also, studies suggest that the classification of childhood cancers should be based on morphology rather than topography. The ICD-O-3 is designed to categorize primarily by the site of origin, which is suitable for representing adult cancers. The International Classification of Childhood Cancer (ICCC), developed under the auspices of the International Agency of Research of Cancer (IARC), the International Association of Cancer Registries, and the Société Internationale d’Oncologie Pédiatrique (SIOP), is designed to emphasize the histology of tumors and leverages a combination of site and histology to characterize and classify childhood cancers. The information extraction from childhood cancer pathology reports should emphasize the morphology rather than the primary site of origin for these cancers.

The present study aims to develop an optimal ML model for automatic information extraction for pediatric cancer pathology reports based on ICCC coding and to establish a high-precision model by applying the uncertainty quantification (UQ) mechanism, which is critical for state cancer registries.

To that end, this article (1) developed a model of automatic information extraction from childhood cancer pathology reports based on ICCC, which—to the best of our knowledge—is the first AI/ML model for pediatric cancers; (2) presents results of a model trained on a large volume (29 206 cases) of pediatric cancer cases from 6 state cancer registries; (3) optimized the model for classifying childhood cancer pathology reports; and (4) describes the model calibration using UQ to support human annotators with high precision.

MATERIALS AND METHODS
Data sources
This study’s data set consisted of unstructured text in pathology reports from 6 state cancer registries: the Kentucky Cancer Registry (KCR), Louisiana Tumor Registry (LTR), New Jersey State Cancer Registry (NJSCR), New Mexico Tumor Registry (NMTR), Seattle Cancer Registry (SCR), and the Utah Cancer Registry (UCR). KCR, LTR, NMTR, SCR, and UCR participate in the National Cancer Institute’s (NCI’s) Surveillance, Epidemiology, and End Results (SEER) program. The study was executed according to the institutional review board protocol DOE000619, approved by the US Department of Energy (DOE) Institutional Review Board on April 6, 2021 (initial approval on September 23, 2016). From the data of millions of e-path reports from the cancer registries, we selected cases in which a cancer patient was diagnosed before they were 20 years old.

The gold standard of the abstraction of information extracted from e-path reports is the Cancer/Tumor/Case (CTC) database, which stores all diagnostic, staging, and treatment information for
Table 1. ICCC (a) main and (b) subgroup codes and definitions based on the ICCC third edition

(a)

| Code | Description                                                                 |
|------|-----------------------------------------------------------------------------|
| 01   | Leukemias, myeloproliferative, and myelodysplastic diseases                  |
| 02   | Lymphomas and reticulendothelial neoplasms                                  |
| 03   | CNS and miscellaneous intracranial and intraspinal neoplasms                |
| 04   | Neuroblastoma and other peripheral nervous cell tumors                      |
| 05   | Retinoblastoma                                                              |
| 06   | Renal tumors                                                                |
| 07   | Hepatic tumors                                                              |
| 08   | Malignant bone tumors                                                       |
| 09   | Soft tissue and other extraosseous sarcomas                                 |
| 10   | Germ cell tumors, trophoblastic tumors, and neoplasms of gonads             |
| 11   | Other malignant epithelial neoplasms and malignant melanomas                |
| 12   | Other and unspecified malignant neoplasms                                   |
| 999  | Not classified by SEER or in situ                                           |

(b)

| Code | Description                                                                 |
|------|-----------------------------------------------------------------------------|
| 014  | Myelodysplastic syndrome and other myeloproliferative diseases              |
| 015  | Unspecified and other unspecified leukemias                                 |
| 021  | Hodgkin lymphomas                                                           |
| 022  | Non-Hodgkin lymphomas                                                       |
| 023  | Burkitt lymphoma                                                            |
| 024  | Miscellaneous lymphoreticular neoplasms                                     |
| 025  | Unspecified lymphomas                                                       |
| 031  | Ependymomas and chorioid plexus tumor                                       |
| 032  | Astrocytomas                                                                |
| 033  | Intracranial and intraspinal embryonal tumors                              |
| 034  | Other gliomas                                                               |
| 035  | Other specified intracranial and intraspinal neoplasms                      |
| 036  | Unspecified intracranial and intraspinal neoplasms                         |
| 041  | Neuroblastoma and ganglioneuroblastoma                                     |
| 042  | Other peripheral nervous cell tumors                                        |
| 050  | Retinoblastoma                                                              |

Table 1. continued

| Code | Description                                                                 |
|------|-----------------------------------------------------------------------------|
| 061  | Nephroblastoma and other nonepithelial renal tumors                         |
| 062  | Renal carcinomas                                                            |
| 063  | Unspecified malignant renal tumors                                          |
| 071  | Hepatoblastoma and mesenchymal tumors of the liver                          |
| 072  | Hepatic carcinomas                                                          |
| 073  | Unspecified malignant hepatic tumors                                        |

The total number of childhood cancer pathology reports in our data corpus is 29,206 from 11,274 patients. Figure 1 illustrates the number of cases per each ICCC code. Leukemias (01) and lymphomas (02) represent more than half of all childhood cancers. Leukemia is the most prevalent cancer in children, and this finding is consistent with existing research. Lymphoid leukemia (011) is the most prevalent leukemia and represents more than 25% of childhood cancer cases. Among lymphomas, non-Hodgkin lymphoma (022) is the most prevalent type. Note that the figure illustrates the severe class imbalance within this data set.

Figure 2 illustrates the number of cases per ICCC code sorted by a patient’s age at diagnosis. Leukemia (01) is more common in younger patients (ages 0–4) but diminishes among older children. Similar patterns occurred for patients with neuroblastoma (04). In contrast, the incidence of lymphomas (02) was highest among adolescents. Germ cell tumors (10) and other malignant epithelial tumors and melanomas (11) are also most common among young adolescents. Note that the number of cancer cases between ages 5 and 11 is considerably lower than for the other age groups. The observations and findings are consistent with the reports and statistics from other studies, which implies that the data set from the 6 population-based registries in SEER included in this study reflects the real-world situation.
ML models for text classification

TextCNN\textsuperscript{4,14} is one of the most successful and widely used convolutional neural network (CNN) models for text comprehension and classification. It consists of 3 parts: word embedding, 1D convolution, and a fully connected decision layer. Word embedding is a learned representation of terms and words to map a set of words onto vectors of numerical representations with the same semantic meaning and similar observation. The 1D convolution layer has a series of 1D convolution filters that have latent representations to articulate the features in the word vectors of documents. The features found are passed to the fully connected layer to make inferences. MT-CNN\textsuperscript{5} extends TextCNN by adding a multitask learning (MTL) mechanism\textsuperscript{15} to the decision layer. A classifier learns multiple tasks simultaneously and finds an optimal latent representation to solve a series of related tasks. The MTL helps find more generalized solutions than single-task models, thus yielding higher task performance. We have successfully developed an MT-CNN model for automatic information extraction based on ICD-O-3 and verified that the CNN model has competitive task performance while exhibiting prompt training and inference time.\textsuperscript{16}

Automatic information extraction based on ICCC

The following subsections describe 2 models that we designed and tested for this study along with 2 scenarios for each model.

Model 1: ICD-O-3 classification then ICCC recoding approach

The first model involves the classification of ICD-O-3. Generally, the ICCC coding is a recoding based on the site, histology, and behavior from the ICD-O-3 codes. Therefore, the automatic information extraction from childhood cancer pathology reports can be accomplished using the existing classification model\textsuperscript{5} for cancer patients of all ages. This approach saves the time and effort required to develop a new model for classifying cancer pathology reports based exclusively on the ICCC coding.

However, some factors may cause a decrease in classification accuracy. First, ICCC includes 47 codes, whereas ICD-O-3 con-
sists of more than 300 site codes and 600 histology codes. Designing and training an ML/DL model with that many labels could be overly complex and prone to error. Second, as stated earlier, certain cancer types are more prevalent in adults than in children and vice versa; moreover, cancer is more prevalent in adults, generally. Consequently, the model trained on the entire corpus of cancer patients could be skewed more toward the reports from adult patients.

We developed 2 scenarios to evaluate if we can achieve better accuracy by limiting the scope to pathology reports of childhood cancers.

1. Model 1(a): ICD-O-3 classification model is trained by cancer pathology reports from all age groups and then recoded for ICCC
2. Model 1(b): ICD-O-3 classification model is trained by the childhood cancer pathology reports only and then recoded for ICCC

Model 2: Direct ICCC classification approach
The second model involves classifying ICCC codes directly from the cancer pathology reports. The new model was trained on ICCC data and may have higher classification accuracy because it only deals with 47 classes, whereas the ICD-O-3-based models must contend with numerous class labels. This approach required us to train and deploy another ML model specifically for childhood cancer pathology reports, which required extra effort and resources.

In addition, we conducted a further study based on the consensus recommendation from the Childhood Cancer Data Initiative’s (CCDI’s) advisory group that the upper age limit of diagnosis be up to 39 years old for certain childhood/pediatric cancers. Table 2 lists the cancer types to be regarded as pediatric cancers at this upper age limit per CCDI’s suggestion. Note that 1055 cases fall into these categories, which is a relatively small number.

We developed 2 scenarios to quantify the effect of augmenting the data per CCDI’s recommendation.

1. Model 2(a): ICCC classification model is trained by the childhood cancer pathology reports
2. Model 2(b): ICCC classification model is trained by the childhood cancer pathology reports with augmentation suggested by CCDI

Figure 3 illustrates the architectures of the 2 models (and 4 scenarios) that we designed and evaluated.

Uncertainty quantification
The purpose of automatic information extraction is to either assist human observers with a second opinion or automate coding where feasible to enable humans to focus on cases that are more complex or challenging. To this end, the most important feature that the model should possess to achieve the objective is a highly reliable and accurate decision from the model. Inaccurate second opinions (from the model) may distract human observers and even degrade the process’ perfor-

| Code | Description | No. of cases |
|------|-------------|--------------|
| 061  | Nephroblastoma and other nonepithelial renal tumors | 26 |
| 071  | Hepatoblastoma and mesenchymal tumors of the liver | 11 |
| 081  | Osteosarcoma | 301 |
| 082  | Chondrosarcoma | 232 |
| 083  | Ewing tumors and related sarcoma of the bone | 193 |
| 084  | Other specified malignant bone tumors | 90 |
| 085  | Unspecified malignant bone tumors | 23 |
| 091  | Rhabdomyosarcomas | 179 |

Figure 2. Number of childhood cancer pathology reports by ICCC main codes and age at diagnosis.

Table 2. List of cancers diagnosed between ages 20 and 39 that could be regarded as pediatric cancers per CCDI and the number of cases that are augmented to the training of Model 2(b)
mance. If the model’s decision is incorrect, then review is needed, which limits the efficiency and benefit of using an automated process. UQ, which is vital to the process, assigns a confidence estimate to the machine-assigned code to allow a human to determine whether additional review is necessary, thus minimizing human labor.

In this article, we propose a post-training threshold approach based on the estimation of a confidence score from the softmax-predicted probabilities in the validation set (rather than the training set).

Let \( Y = (y_1, \ldots, y_n) \) be the softmax-predicted probabilities for the \( n \) classification labels. The confidence score is estimated as the conditional distribution of a correct classification via the Bayes theorem for a binary variable (in which being correct and incorrect are mutually exclusive outcomes):

\[
p(\text{correct}|y) = \frac{p(y|\text{correct}) \cdot p(\text{correct})}{p(y|\text{correct}) \cdot p(\text{correct}) + p(y|\text{incorrect}) \cdot p(\text{incorrect})}
\]

The marginal probabilities \( p(\text{correct}) \) and \( p(\text{incorrect}) \) are called priors and are estimated as the corresponding relative frequencies (ie, the total number of correct or incorrect decisions divided by the total number of cases).

The conditional probabilities \( p(y|\text{correct}) \) and \( p(y|\text{incorrect}) \) are estimated from the data by modeling the conditional quantile functions \( F^{-1}_\text{correct}(\tau|\text{correct}) \) and \( 1 - F^{-1}_\text{incorrect}(\tau|\text{incorrect}) \) for a selected percentile point, \( \tau \).

**Experimental setup**

We designed a comparison study to determine the classification accuracies of the models described above. F1 scores, a widely accepted metric for information retrieval, are used as the performance benchmark. Because of the severe class imbalance of the data corpus, we employed both micro-averaged and macro-averaged F1 scores. The micro-F1 is weighted equally to the individual cases, whereas the macro-F1 is weighted equally to the class label. If the ML model was favorable to the prevalent class labels but did not work well with the samples from minor classes, then the macro-F1 score would be lower than the micro-F1 score.
currence of “ultra-rare” pediatric tumors. However, the severity of this imbalance is substantially smaller than for the ICD-O-3 system as applied to pediatric cancers.

Performance degradation in information extraction models caused by underrepresented class labels in the data corpus is a critical issue in developing algorithms for automation in cancer surveillance. There is no definitive way to increase the sample size of cancer pathology reports from rare cancers given the rarity of pediatric tumors in general (16,000 cases in the United States per year). One might suggest special ML techniques, such as data synthesis, but it is well known that synthesizing free-form text data is not a trivial task.

One possible solution is to augment the training corpus with the reports from the subjects of age 20–39, as some ICCC sites are considered pediatric tumors even when occurring in adolescents or young adults. The results listed in Table 3 show the effects of adding those reports based on expert consensus in the CCDI community. The class labels that already had more than 100 samples in the corpus did not benefit from the augmentation because the model had already achieved high classification accuracy scores for those labels. However, substantial improvements were made for code 084—Other specified malignant bone tumors (from 0.76–F1 to 0.85–F1) and for code 085—Unspecified malignant bone tumors (from 0.38–F1 to 0.60–F1). Those 2 labels had a relatively small number of samples in the corpus: 30 cases for code 084 and 25 cases for code 085. Adding 90 cases for code 084 and 23 cases for code 085 increased the chance of learning features for correct decisions for those class labels.

Note, however, that Model 1(b) recorded higher accuracy scores than Model 1(a), which implies that simply adding adult cancer pathology reports did not improve accuracy. This makes sense because the corpus contained many more adult cancers than childhood cancers, and the prevalent adult cancers (eg, breast cancer) are rare in childhood. Thus, this simple addition may not improve the ML model’s performance for childhood cancer cases, which leads to the following question: can we improve the accuracy score if we curate the augmented data set by undersampling the breast and lung cancers or by downselecting the cancer types that are more prevalent in children?

This would make an excellent future research topic. This solution of expanding the corpus based on codes, combined with expanding the training corpus by bringing in additional registries, might serve as a partial but nontrivial solution to reduce the class imbalance.

The application of the softmax-based UQ mechanism was successful. We demonstrated that Model 2(b) with the softmax-based UQ achieved 0.987—a nearly perfect micro-F1 score—while it abstained in only 14.8% of the cases. This result implies that the system could serve human annotators at state cancer registries. The model can process more than 85.2% of the childhood cancer pathology reports with confidence. Further manual review is needed for only 14.8% of the incoming data, which indicates that the model is highly reliable and potentially ready to use.

Several factors contribute to abstention of cases, such as the case difficulty, incompleteness of information on the data samples, or not enough information supplied to the classification model owing to a lack of training samples. One clue is that the abstention rate (ie, the number of abstained cases divided by the total number of cases in the class label) is considerably lower for the prevalent classes. For example, the abstention rate of code 011 was 0.108 (871/8042), whereas the abstention rate for code 014 was 0.535 (76/137). Increasing performance for those underrepresented classes is key to achieving a more reliable model.

Some questions remain. The choice of 0.9 as the threshold of probability for correct decision-making is entirely arbitrary and unlikely to be sufficiently robust to meet the high-quality standards of the surveillance community. The threshold value is subject to the tolerance level of wrong decisions by the state cancer registries and will determine the credibility of the data products from the registries. Follow-up statistical analyses of this threshold are required.

CONCLUSION

In this article, we described our study of the classification of childhood cancer pathology reports in terms of the ICCC coding and established an automatic information extraction system for processing a massive volume of pathology reports suitable for state cancer registries. We tested the 2 models: (1) reusing the existing model for extracting ICD-O-3 codes and recoding them into ICCC and (2) developing a new model for extracting ICCC coding. We also experimented with a softmax-based UQ algorithm to evaluate model performance when discarding the minimum amount of ambiguous cases. Our findings suggest that the model for classifying ICCC coding with UQ is suitable for alleviating the workload of human annotators at state cancer registries.

FUNDING

This work was supported in part by the Joint Design of Advanced Computing Solutions for Cancer (JDACS4C) program established by DOE and the NCI of the National Institutes of Health. This work was performed under the auspices of DOE by Argonne National Laboratory under Contract DE-AC02-06-CH11357, Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344, Los Alamos National Laboratory under Contract DE-AC5206NA25396, and Oak Ridge National Laboratory under Contract DE-AC05-00OR22725.

AUTHOR CONTRIBUTIONS

H-JY carried out the problem conception, implementation, validation tests, and drafted manuscript. AP carried out the design of uncertainty quantification algorithm. EBD, X-CW, AS, JD, SS, CW, and LC carried out the data curation. EBD, X-CW, and LP helped to draft the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

This manuscript has been authored in part by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

The authors would like to acknowledge the contribution to this study from other staff in the participating central cancer registries. These registries are supported by the National Cancer Institute’s Surveillance, Epidemiology, and End Results (SEER) Program, the Centers for Disease Control and Prevention’s National Program of Cancer Registries (NPCR), and/or state agencies, universities, and cancer centers. The participating central cancer registries include the following: Kentucky working under contract numbers SEER: HHSN261201800013I/HHSN261100001 and NPCR:
CONFLICT OF INTEREST STATEMENT
No competing interests to declare.

DATA AVAILABILITY
The data underlying this article were provided by the state cancer registries participated in the study by permission and cannot be shared publicly due to the privacy of individuals in the data corpus.

REFERENCES
1. Siegel RL, Miller KD, Fuchs HE, et al. Cancer statistics, 2021. CA Cancer J Clin 2021; 71 (1): 7–33.
2. Ward E, DeSantis C, Robbins A, et al. Childhood and adolescent cancer statistics, 2014. CA Cancer J Clin 2014; 64 (2): 83–103.
3. Steliarova-Foucher E, Colombet M, Ries LA, et al. International incidence of childhood cancer, 2001–10: a population-based registry study. Lancet Oncol 2017; 18 (6): 719–31.
4. Qiu JX, Yoon HJ, Fearn PA, et al. Deep learning for automated extraction of primary sites from cancer pathology reports. IEEE J Biomed Health Inform 2018; 22 (1): 244–51.
5. Alawad M, Gao S, Qiu JX, et al. Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks. J Am Med Inform Assoc 2020; 27 (1): 89–98.
6. Gao S, Qiu JX, Alawad M, et al. Classifying cancer pathology reports with hierarchical self-attention networks. Artif Intell Med 2019; 101: 101726.
7. Gao S, Alawad M, Young MT, et al. Limitations of transformers on clinical text classification. IEEE J Biomed Health Inform 2021; 25 (9): 3596–607.
8. American Cancer Society. Cancer Statistics Center. http://cancerstatistics-center.cancer.org. Accessed June 9, 2022.
9. Kramárová E, Stiller CA. The international classification of childhood cancer. Int J Cancer 1996; 68 (6): 759–65.
10. Steliarova-Foucher E, Stiller C, Lacour B, Kaatsch P. International classification of childhood cancer. Cancer 2005; 103 (7): 1457–67.
11. International Classification of Childhood Cancers (ICCC). https://seer.cancer.gov/iccc. Accessed June 9, 2022.
12. Miller RW, I. Young J Jr, Novakovic B. Childhood cancer. Cancer 1995; 75 (51): 395–405.
13. Cancer Incidence Statistics. https://www.cancerresearchuk.org-health-professional/cancer-statistics/incidence#heading-Two. Accessed June 9, 2022.
14. Kim Y. Convolutional neural networks for sentence classification. arXiv:14085882. 2014.
15. Yoon HJ, Ramanathan A, Tourassi G. Multi-task deep neural networks for automated extraction of primary site and laterality information from cancer pathology reports. In: INNS Conference on Big Data. Springer; 2016: 195–204.
16. Qiu JX, Yoon H-J, Srivastava K, et al. Scalable deep text comprehension for cancer surveillance on high-performance computing. BMC Bioinformatics 2018; 19 (S18): 99–110.
17. Fritz AG. International Classification of Diseases for Oncology: ICD-O. World Health Organization; 2000. https://apps.who.int/tris/handle/10665/42344. Accessed June 9, 2022.
18. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011; 12: 2825–30.
19. Chollet F, et al. Keras. 2015. https://github.com/fchollet/keras. Accessed June 9, 2022.
20. Abadi M, Agarwal A, Barham P, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015. http://tensorflow.org/. Accessed June 9, 2022.