A Survey on Publicly Available Open Datasets Derived From Electronic Health Records (EHRs) of Patients with Neuroblastoma

REVIEW

DAVIDE CHICCO
GABRIEL CERONO
DAVIDE CANGELOSI

*Author affiliations can be found in the back matter of this article

ABSTRACT

Background: Neuroblastoma is a rare pediatric cancer that affects thousands of children worldwide. Information stored in electronic health records can be a useful source of data for *in silico* scientific studies about this disease, carried out both by humans and by computational machines. Several open datasets derived from electronic health records of anonymized patients diagnosed with neuroblastoma are available in the internet, but they were released on different websites or as supplementary information of peer-reviewed scientific publications, making them difficult to find.

Methods: To solve this problem, we present here this survey of five open public datasets derived from electronic health records of patients diagnosed with neuroblastoma, all collected in a single website called Neuroblastoma Electronic Health Records Open Data Repository.

Results: The five open datasets presented in this survey can be used by researchers worldwide who want to carry on scientific studies on neuroblastoma, including machine learning and computational statistics analyses.

Conclusions: We believe our survey and our open data resource can have a strong impact in oncology research, allowing new scientific discoveries that can improve our understanding of neuroblastoma and therefore improve the conditions of patients. We release the five open datasets reviewed here publicly and freely on our Neuroblastoma Electronic Health Records Open Data Repository under the CC BY 4.0 license at:

https://davidechicco.github.io/neuroblastoma_EHRs_data or at

https://doi.org/10.5281/zenodo.6915403

CORRESPONDING AUTHOR:

Davide Chicco

Institute of Health Policy Management and Evaluation, University of Toronto, Toronto, Ontario, Canada

davide.chicco@gmail.com

KEYWORDS:

neuroblastoma; electronic health records; EHR; datasets; data; open data; neuro-oncology; childhood cancer; pediatric oncology

TO CITE THIS ARTICLE:

Chicco, D, Cerono, G and Cangelosi, D. 2022. A Survey on Publicly Available Open Datasets Derived From Electronic Health Records (EHRs) of Patients with Neuroblastoma. *Data Science Journal*, 21: 17, pp. 1–18. DOI: https://doi.org/10.5334/dsj-2022-017

1 INTRODUCTION

Neuroblastoma is a pediatric cancer that forms in certain types of nerve tissue, most commonly arising from the adrenal glands. Neuroblastoma is the most common cancer among newborns and affects thousands of children worldwide (American Cancer Society, 2021). Ninety percent of neuroblastoma cases are diagnosed by age five. Neuroblastoma forms in specific types of nerve tissue and usually develops in the adrenal glands in the abdomen (Colon & Chung 2011). The 5-year survival rate for children with neuroblastoma in England is approximately 67% (Children with cancer UK, 2021). High-risk neuroblastoma is usually treated with intensive chemotherapy, surgery, radiation therapy, bone marrow and hematopoietic stem cell transplantation.

Knowledge about this pediatric cancer can be discovered not only with laboratory results and clinical trials, but also by analyzing the data contained in the electronic health records (EHRs) of patients. Computational statistical methods and machine learning techniques applied to data derived from structured EHRs can, in fact, be effective tools to infer new knowledge about neuroblastoma and eventually to help medical doctors to develop better treatments (Adkins 2017).

Datasets derived from electronic health records, although extremely useful for the progress of scientific research, are often kept reserved and unshared with the rest of the scientific community, because of privacy issues or because of lack of data sharing culture in the hospital or research centre. In recent years, a wave of requests for open data sharing has been launched by researchers around the world, which culminated in the definition of FAIR principles for data sharing (Bertagnolli et al. 2017; Stall et al. 2019; Wilkinson et al. 2016).

The advantages of open data sharing have been already demonstrated by important initiatives that have had a profound impact on scientific research. The University of California Irvine Machine Learning Repository (University of California Irvine 1987), for example, is an online catalogue of free datasets coming from several domains (biology, medicine, physics, computer science and engineering, social sciences, business, and games) that started in 1987 and now contains 588 different datasets. It includes some derived from electronic health records, but not of neuroblastoma.

Another online resource that shares public datasets is Kaggle. Since 2010, this online community of data scientists has collected and provided open datasets to their users, to be used for scientific competitions or for independent analyses. To date, Kaggle lists around ten thousand public datasets (Kaggle 2022). In 2017, Google launched the Dataset Search (Google 2022) engine, where users can easily look for public datasets on the internet. Re3data (2022) is another interesting resource to mention: it a public registry of scientific data repositories available online.

In bioinformatics, Gene Expression Omnibus (GEO) (US National Center for Biotechnology Information 2021) has been providing thousands of open gene expression and methylation datasets to researchers worldwide, producing thousands of studies and publications too. GEO contains gene expression data of neuroblastoma, too, that led to some significant genetic discoveries (Cangelosi et al. 2020; Melaiu et al. 2020). Neuroblastoma bioinformatics data were publically shared and integrated for the CAMDA 2017 conference Neuroblastoma Data Integration Challenge (Francescatto et al. 2018; CAMDA 2017). Multiple datasets of images have also recently been released on public repositories: ophthalmological images (Khan et al. 2021) and cancer images (Clark et al. 2013) among them.

Regarding neuroblastoma, the International Neuroblastoma Risk Group (INRG) recently released the INRG Data Commons (International Neuroblastoma Risk Group 2017; Volchenboum et al. 2017), a database of thousands of EHRs of patients with neuroblastoma. Despite its usefulness, the access to the INRG Data Commons is restricted to the participants to approved projects: to obtain the data, one needs to fill and submit a project application, which might or might not be approved by the INRG Data Commons board. Even in the case of proposal acceptance, this procedure clearly requires some processing time: between the proposal submission and the actual data download, months or even years can go by, with consequent delays that can negatively affect the project itself. Of course, this restricted approach limits the possibility of using their data.

A similar initiative, the Pediatric Cancer Data Commons (Plana et al. 2021; Volchenboum et al. 2021), was launched by University of Chicago and provides EHRs data of patients with

childhood cancers, including neuroblastoma. However, the access to this dataset is restricted to pre-authorized researchers, too. These limitations of access generate several problems: pre-authorization can limit data access and research transparency and sharing of work, that could be the basis for more research.

Since a public free resource containing public open unrestricted data derived from electronic health records of patients diagnosed with neuroblastoma is currently missing, we gathered the five open datasets derived from EHRs of neuroblastoma described in this survey in a website called Neuroblastoma Electronic Health Records Open Data Repository, that we freely released online. Moreover, we converted these datasets into numerical values, making them computer-readable for any computational analyses.

Users and researchers can take advantage of this resource by downloading the datasets and performing their scientific analyses that might lead to new discoveries about this pediatric cancer. Moreover, we performed a detailed descriptive analysis of all the clinical features present in the five listed datasets, by providing detailed information about all the variables. Some information about these variables were absent even in the original datasets publications.

We organize the rest of this study as follows. After this Introduction, we describe the datasets of our survey in section 2, and we discuss the main insights of their clinical variables in section 3. We finally outline some conclusions in section 4.

2 DATA AND METHODOLOGY

We performed a thorough literature search of scientific articles related to neuroblastoma electronic health records (EHRs) in November and December 2020, using the Google Scholar (Google 2021) search engine. We performed this dataset search by following the guidelines by Marco Pautasso (Pautasso 2013), by looking for the keywords 'electronic medical records and neuroblastoma', 'EMR and Neuroblatoma', 'electronic health records and Neuroblastoma', 'EHR and neuroblastoma', and 'clinical records and neuroblastoma'. Following the example of Khan and colleagues (2021), we collated and screened the results returned from the first ten pages of the search.

We identified nine articles including EHRs data of patients diagnosed with neuroblastoma in their main texts or in their supplementary information.

Five of them have public open datasets released under CC BY 4.0 license and therefore downloadable and usable without non-commercial restrictions (Banelli et al. 2013; Kim et al., 2018; Villamón et al., 2013; Choi et al., 2019; Ma et al., 2018), included in the supplementary information of the articles. Three of them contain datasets, but without an open license, and therefore cannot be used (Federico et al. 2015; Rosenbaum et al. 2013; Smith et al. 2010). Moreover, one article contains an open dataset, but it is related to coronary artery disease and only minimally pertinent to neuroblastoma (Matsumura et al. 2018).

We report the quantitative characteristics of these five open datasets in Table 1. Each of these five open datasets is available as an CSV file, thus it can be opened with most spreadsheet software, including free open source available ones such as LibreOffice Calc (The Document Foundation 2022) or processed with appropriate software packages on any computer, in `R` for example (Software Carpentry 2022).

DATASET NAME	REFERENCE	#PATIENTS	#FEATURES	#MISSING VALUES	TABLE
dataBB2013	Banelli et al. (2013)	121	11	50	Table 2
dataCK2018	Kim et al. (2018)	20	16	29	Table 3
dataEV2013	Villamón et al. (2013)	19	11	13	Table 4
dataYBC2019	Choi et al. (2019)	7	10	0	Table 5
dataYM2018	Ma et al. (2018)	169	13	52	Table 6
interval		[7, 169]	[10, 16]	[0, 52]	
mean		67.2	12.2	28.8	

Table 1 Quantitative characteristics of the analyzed datasets. #patients: number of patients. #features: number of clinical features. #missing values: number of missing data instances.

We generated the frequency table image Figure 1 with the `ggplot2` R package (Wickham, 2016).

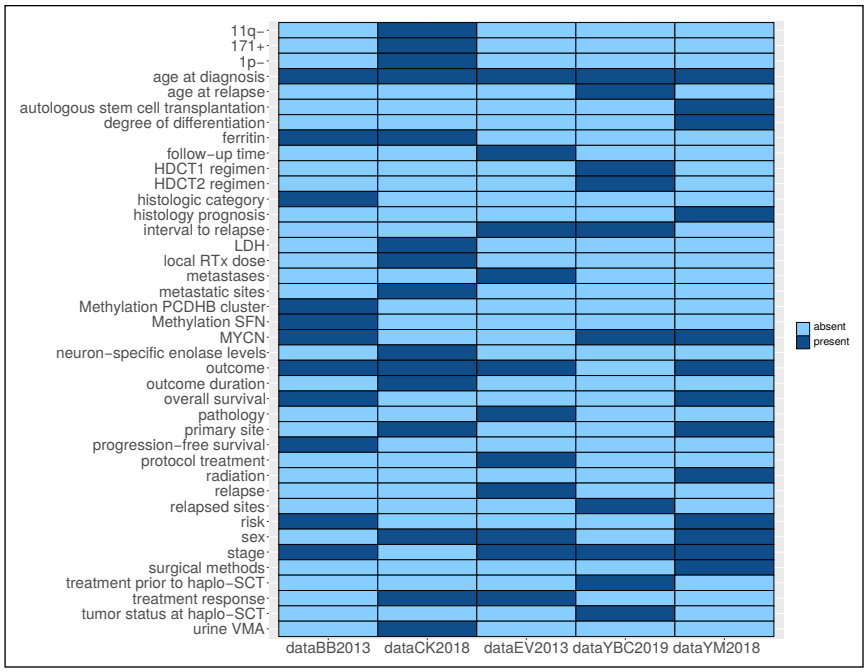

Figure 1 Presence and absence of the clinical features in the five analyzed datasets. x axis: datasets names. y axis: clinical features of the datasets. dataBB2013: Banelli et al. (2013) (Table 2). dataCK2018: Kim et al. (2018) (Table 3). dataEV2013: Villamón et al. (2013) (Table 4). dataYBC2019: Choi et al. (2019) (Table 5). dataYM2018: Ma et al. (2018) (Table 6).

2.1 DETAILS ON THE DATASETS

In this section, we describe the details about the datasets, including the quantitative characteristics of the patients and of the clinical features, and the meaning of each of them.

2.1.1 dataBB2013

The dataBB2013 dataset was released by Banelli et al. (2013), and was previously employed by the same research team in earlier studies (Banelli et al. 2005b,a, 2010). The dataset contains data from 121 patients, including survived and deceased subjects, collected at the hospital of Istituto Giannina Gaslini (Genoa, Italy) between 1990 and 2004. Each patient has clinical features (Table 2), but one of them is fixed (4th Stage of International Neuroblastoma Staging System – INSS).

DATABB2013			
FEATURE	**MEANING**	**TYPE**	**VALUES**
age at diagnosis	age at diagnosis	integer	2, ..., 196
ferritin	ferritin serum level	ng/ml	−99, 19, ..., 2250
histological category	histological category of the neuroblastoma	categorical	NS, NB, GNB
INRG Risk classification	risk group: HR high risk and I/LR intermediate/low risk	categorical	0, 1
INSS Stage	stage of the tumor (only stage 4 patients)	categorical	1
MYCN amplification	status of nMYC oncogene: 0, amplified, 1, unamplified;	binary	0, 1
Methylation PCDHB cluster (%)	methylation of 17 genes of the Protocadherin B cluster	percentage	29.44, ..., 88.93
Methylation SFN (%)	methylation of the SFN gene	percentage	36.6, ..., 99
OS	overall survival	float	0.27, ..., 164.47
outcome	clinical outcome	categorical	AICR, AWD,
			AWSD, CR, DOD
PFS	progression free survival	float	3.7, ..., 74.7, NaN

Table 2 Meaning and values of the features of the dataBB2013 dataset. Number of patients: 121. Number of features: 11. AICR: alive in complete remission. AWD: alive with disease. AWSD: alive with stable disease. CR: complete remission. DOD: Dead of disease. GNB: ganglioneuroblastoma. HR: high risk. INRG: International Neuroblastoma Risk Group MYCN: MYCN oncogene. NaN: not a number. NB: neuroblastoma. NS: not specified, it was impossible to make a more precise diagnosis (Romani, 2021). OS: overall survival. PCDDHB: Protocadherin Beta Cluster. PFS: progression-free survival. SFN: Stratifin gene. Additional information can be found in the dataset original article by Banelli et al. (2013).

The dataset includes some typical features, such as age at diagnosis, histological category, risk group, MYCN amplification, overall survival, progression-free survival, and outcome. The outcome indicates if the patient survived or not. This dataset also contains two features about the methylation of a gene family (Protocadherin Beta Cluster, PCDHB) and of a specific gene (SFN). The prognostic role of the methylation of these two genes is the key aspect of the original study (Banelli et al. 2013).

The dataBB2013 dataset is the only one including data related to the methylation of the PCDHB gene family and to the methylation of the SFN gene. The methylation of Protocadherin Beta cluster genes is known to be associated to poor prognosis in patients with neuroblastoma (Abe et al. 2005), while the association between the methylation of the protein-coding Stratifin gene and advanced stage, high-risk neuroblastoma was shown by the dataset authors in a previous study (Banelli et al. 2010).

This dataset contains only one blood test variable: ferritin. Ferritin is a protein that contains iron and is known to be a clear prognostic factor for neuroblastoma (Moroz et al. 2020).

2.1.2 dataCK2018

The dataCK2018 dataset was published by Kim et al. (2018). It contains data of 20 patients, collected at the Samsung Medical Center from January 2009 to December 2015 at the Samsung Medical Centre of the Sungkyunkwan University (Seoul, South Korea). All the patients had chemotherapy and surgery, with or without local radiation therapy, followed by differentiation therapy. Each patient has 16 clinical factors (Table 3), making it the dataset with the higher number of features among the five listed in this article.

DATACK2018			
FEATURE	**MEANING**	**TYPE**	**VALUES**
11q–	presence of chromosomal aberration at 11q site	Boolean	yes, no, –
17q+	presence of chromosomal aberration at 17q site	Boolean	yes, no, –
1p–	presence of chromosomal aberration at 1p site	boolean	yes, no, –
age	age at diagnosis	months	0, ..., 10.2
ferritin	ferritin levels	ng/mL	15, ..., 1638.6
LDH	lactic acid dehydrogenase level	U/L	539, ..., 6200
local RTx dose (gy)	dosage of radiation	float	15, 23.4, 25.2, –
			BM, bone, kidney, liver, LN,
metastatic sites	site of metastsoization	categorical	lung, mediastinum, muscle,
			pleura, skin
NSE	neuron-specific enolase levels	ng/mL	7.3, ..., 947
outcome	event-free survival (EFS) or No Evidence of Disease	Boolean	EFS, NED
outcome_mo	follow-up	months	17, ..., 91
primary site	tumor primary site	binary	abdomen, mediastinum
sex	male or female	binary	M, F
tumor response after CT & S	response after chemotherapy and surgery	categorical	CR, MR, PR, VGPR
tumor response after DT	response after differentiation theraphy (DT)	categorical	CR, MR, PR, VGPR
urine VMA	vanillylmandelic acid levels in urine	mg/day	0.5, ..., 53.9

Table 3 Meaning and values of the features of the dataCK2018 dataset. Number of patients: 20. Number of features: 16 BM: bone marrow. CR: complete response. CT: chemotherapy. DT: differentiation therapy. F: female. gy: gray units. ID: identifier. LDH: lactic acid dehydrogenase. LN: lymph node. LSE: neuron-specific enolase. M: male. MR: mixed response. PR: partial response. S: surgery. U/L: units per liter. VGPR: very good partial response. VMA: vanillylmandelic acid. mg: milligrams. mo: months ng/mL: nanograms per milliliter. no: number. In the original article dataset, sex and age are joined in a unique feature called 'Sex/age (mo)', and outcome and outcome months are joined in a unique feature called 'Outcome (mo)'. Additional information can be found in the dataset original article by Kim et al. (2018).

The clinical feature list includes several traditional variables such as age, outcome, sex, follow-up duration, tumor primary site, metastatic sites. Some uncommon variables related to the treatment received by the patients are present: local dosage of radiation, response after chemotherapy and surgery, and response after differentiation therapy.

The main scientific statement of the original study by the authors was that: 'patients younger than 18 months with stage 4 MYCN nonamplified neuroblastoma had high survival changes, without significant late adverse effects, when treated with alternating cycles of cyclophosphamide (CEDC) and ifosfamide, carboplatin, and etoposide (ICE), followed by surgery and differentiation therapy' (Kim et al. 2018).

This dataset contains variables of ferritin and LDH levels in the blood, both recognized as prognostic factors for neuroblastoma in the medical community (Moroz et al. 2020). It includes a variable related to a urine test (urine vanillylmandelic acid), which is an indicator for neuroblastoma diagnosis in children screen tests, and a variable of the blood neuron-specific enolase level, which can be used auxiliary test for neuroblastoma.

This dataCK2018 dataset has the only cohort for which data of genetic factors are indicated: 11q–, 171+, and 1p– chromosomal aberrations. These genetic abnormalities are common in patients diagnosed with neuroblastoma.

2.1.3 dataEV2013

The dataEV2013dataset was made open by Villamón et al. (2013). It contains data from 19 patients and 11 clinical variables (Table 4).

DATAEV2013			
FEATURE	**MEANING**	**TYPE**	**VALUES**
age at diagnosis	age	months	9, ..., 108
follow-up time	overall survival	months	1, ..., 132
metastases	presence of metastasis	boolean	yes, no
outcome	clinical outcome	categorical	ADF, AWD, DOD, DOS, DTC
pathology	pathological category	categorical	nGNB, pdNB, uNB
protocol treatment	treatment protocol	categorical	HR-NBL1, INES, LNESG1, N-II-92, NAR-99
relapse	if the cancer relapsed or not	boolean	yes, no
sex	male or female	binary	M, F
stage	stage of the tumor	categorical	1, 2, 3, 4
time to first relapse	time to first relapse	months	4, ..., 28
treatment response	response to first line treatment	categorical	CR, DP, PR, SurPR, VGPR

Table 4 Meaning and values of the features of the dataEV2013 dataset. Number of patients: 19. Number of features: 11. ADF: alive disease-free. AWD: alive with disease. B: bone. BM: bone marrow. CR: complete response. DOD: died of disease. DOS: died of sepsis. DP: disease progression. DTC: died of treatment complication. F: female. HR-NBL1: High-Risk Neuroblastoma Study 1. INES: Infants Neuroblastoma European Study, SIOPEN protocols. LN: lymph nodes. M: male. N-II-92 and NAR-99: names of national clinical trials in Spain (Noguera 2021). PR: partial response. ST: soft tissue. SurPR: surgical partial resection. VGPR: very good partial response. nGNB: nodular ganglioneuroblastoma. pdNB: poorly differentiated NB. uNB: undifferentiated neuroblastoma. VGPR: very good partial response. Additional information can be found in the dataset original article by Villamón et al. (2013).

All the 11 features are health-related, and they do not contain any genomics or genetic information. The patients' samples were collected between 1999 and 2007, at the Spanish Reference Centre for Neuroblastoma Biological and Pathological studies at the time of diagnosis.

The original study is focused on genetic instability and intratumoral heterogeneity with MYCN amplification, and 11q deletion. According to the NIH Genetic and Rare Diseases Information Center, the chromosome 11q deletion is "a chromosome abnormality that occurs when there is a missing (deleted) copy of genetic material on the long arm (q) of chromosome 11" (National Health Institutes (NIH), Genetic and Rare Diseases Information Center (GARD), 2021), that can cause developmental delay, intellectual disability, behavioral problems and distinctive facial features.

2.1.4 dataYBC2019

The dataYBC2019 dataset was publihsed by Choi et al. (2019). It contains data from 7 patients, each of them with 10 health indicators (Table 5). This dataset has the smallest cohort of patients

among the datasets described in this report, but it is the only complete dataset without any missing values or empty ones. The data of the patients were collected from January 2012 to December 2014 in South Korea.

DATAYBC2019			
FEATURE	**MEANING**	**TYPE**	**VALUES**
age at Dx.	age at diagnosis	years	1.5, ..., 3.5
age at relapse	age at relapse	years	4.1, ..., 8.6
HDCT1 regimen	first high-dose chemotherapy	binary	TTC, CEC
HDCT2 regimen	second high-dose chemotherapy	binary	MEC, MIBG-TM
interval to relapse	interval to relapse	months	12, ..., 75
MYCN status	amplified (A) or not amplified (NA)	binary	A, NA
relapsed sites	relapse sites in the body	categorical	Primary, Brain, Bone, LNs, BM
stage at Dx	only metastistic tumors	categorical	4
treatment prior to haplo-SCT	treatment prior to haploidentical SCT	categorical	Surgery, L-RT, CT×5, CT×6, CT×7
tumor status at haplo-SCT	tumor status at haploidentical SCT	categorical	PR, CR, VGPR

Table 5 Meaning and values of the features of the dataYBC2019 dataset. Number of patients: 7. Number of features: 10. A: amplified. BM: bone marrow. CEC: carboplatin, etoposide, and cyclophosphamide CR: complete response. CT×5: five cycles of chemotherapy. CT×6: six cycles of chemotherapy. CT×7: seven cycles of chemotherapy. Dx: diagnosis. HDCT: high-dose chemotherapy. L-RT: local radiotheraphy. LNs: lymph nodes. MEC: melphalan, carboplatin, and etoposide. MIBG-TM: high-dose 131I-metaiodobenzylguanidine treatment, thiotepa, and melphalan NA: not amplified. PR: partial response. SCT: stem cell transplantation. TTC: topotecan, thiotepa, and carboplatin. VGPR: very good partial response. m: months. y: years. Additional information can be found in the dataset original article by Choi et al. (2019).

The clinical variables of this dataset include traditional factors such as age at diagnosis, MYCN status, as well as uncommon features related to the treatment (HDCT1 and HDCT2 chemotherapy, treatment prior to haploidentical stem cell transplantation). Moreover, this dataset contains multiple interesting factors related to cancer relapse: age at relapse, body sites of the cancer relapse, and interval to relapse. This dataset, however, includes no genomics or genetic variable about the patients' biological profile.

The original study of the dataset curators was focused on verifying if early natural killer cell infusion following haploidentical stem cell transplantation would have reduced the relapse in patients with neuroblastoma (Choi et al. 2019). In both that study and its dataset, the relapse information was of great importance.

2.1.5 dataYM2018

The dataYM2018 dataset was released by Ma et al. (2018). It contains data from 169 patients, with 13 clinical factors (Table 6): it is the largest patients' cohort among the datasets listed in this study. The data were collected at Children's Hospital of Fudan University (China) between 2010 and 2015 (Ma et al. 2018).

This dataset includes several traditional variables that can be found in many neuroblastoma electronic health records, such as age, MYCN status, outcome, sex, risk, and overall survival time. It also includes information about treatment, that is if the patient had autologous stem cell transplantation and radiation, and information about the primary tumor site.

The focus of the original study is the MYCN amplification. The authors claim that the MYCN amplfication was an independently adverse prognostic factor in this cohort of patients with neuroblastoma (Ma et al., 2018).

The autologous stem cell transplantation is one of the therapies employed for patients with neuroblastoma (Trahair et al. 2007).

2.2 SHARED CLINICAL FEATURES

Most frequent features. The age at diagnosis is the only variable present in all the five datasets, while the outcome is present in four out of five (all the datasets except dataYBC2019). The neuroblastoma stage is present in in four out of five datasets, too: all the datasets except

dataCK2018. MYCN and sex are present three times: MYCN in dataBB2013, dataYBC2019, and dataYM2018, while sex in dataCK2018, dataEV2013, and dataYM2018. Seven other variables are present in two datasets (ferritin, histology, interval to relapse, overall survival, primary site, risk, treatment response). All the other features are present only in one dataset.

Table 6 Meaning and values of the features of the dataYM2018 dataset. Number of patients: 169. Number of features: 13. FH: favorable histology. MYCN: MYCN oncogene. UH: unfavorable histology. Additional information can be found in the dataset original article by Ma et al. (2018).

DATAYM2018			
FEATURE	**MEANING**	**TYPE**	**VALUES**
age	0: < 12 months; 1: 12–60 months; 2: ≥ 60 months.	integer	0, 1, 2
autologous stem	autologous stem cell transplantation: 0: no; 1: yes.	binary	0, 1
cell transplantation			
degree of differentiation	0: undifferentiated; 1: poorly differentiated; 2: differentiated.	categorical	0, 1, 2
histology prognosis	1: FH favorable histology, 0: UF unfavorable histology	binary	0, 1
MYCN status	status of nMYC oncogene: 0: amplified; 1: unamplified.	binary	0, 1
outcome	clinical outcome: 1, dead of disease, 0, alive or lost follow-up.	binary	0, 1
radiation	if the patient had radiation	boolean	0, 1
risk	risk group: 0: intermediate-risk; 1: high-risk.	categorical	0, 1
sex	0: male; 1: female.	integer	0, 1
site	primary tumor site: 0: adrenal gland; 1: mediastinum; 2: others.	categorical	0, 1, 2
stage	stage of the tumor	categorical	1, 2, 3, 4
surgical methods	total or partial resection	binary	0, 1
time	overall survival	months	1, ..., 100

We show the features present in each dataset in Figure 1.

INRG features. Some clinical variables refer to the International Neuroblastoma Risk Group (INRG) classification system, which has been developed to establish an international consensus for pre-treatment risk stratification and addressing patients to the most suitable treatment protocol (Cohn et al. 2009). Age at diagnosis, tumor stage, MYCN status, histologic category, degree of tumor differentiation, ploidy, and loss of 11q (11q–) are currently used in the INRG schema to assign a very-low, low, intermediate or high-risk group. schema (Cohn et al. 2009). All features, except ploidy, have been reported in at least one of the studies (Figure 1).

Treatment variables. In a modern therapy, heterogeneous treatment ranging from observation, for very low-risk patients, to combinations of intensive multi-agent induction chemotherapy, surgery, radiation, myeloablative consolidation therapy with stem cell rescue and transplantation, 13-cis retinoic acid, and immunotherapy for high-risk patients are provided to patients diagnosed with neuroblastoma (Qi & Zhan 2021). Risk group, treatment protocol, surgical method, radiation, autologous stem cell transplantation, or local radiation theraphy (RTx) dose were reported in one of the studies (Figure 1). Standardized methods to define and interpret response to first line (diagnosis) are important to efficiently monitoring and advancing therapy for neuroblastoma (Park et al. 2017). Response to first line treatment has been included in three out five studies, but this clinical feature has been indicated with different names including *tumor response* (Kim et al. 2018), *treatment response* (Villamón et al. 2013), and *tumor status at haplo-SCT* (Choi et al. 2019).

Sex. Significant positive, but modest, association has been observed between male sex and neuroblastoma (Williams et al. 2019). Although a significant impact of sex to neuroblastoma prognosis or diagnosis has not been described yet, the sex feature was included in four out five datasets (Figure 1). Serum lactate dehydrogenase (LDH), serum ferritin, neuron-specific enolase (NSE), urine vanillylmandelic acid (VMA) are well-known catecholamines used in the

clinical setting to perform neuroblastoma diagnostic and prognostic evaluations (Barco et al. 2014; Cangemi et al. 2012; Ferraro et al. 2020; Tolbert & Matthay 2018).

LDH, ferritin, genetic aberrations. Serum lactate dehydrogenase (LDH) and serum ferritin are strong prognostic molecular markers with potential ability to identify ultra-high-risk and to refine risk stratification (Moroz et al. 2020). Although the clinical utility of these molecular markers has been reported in the literature, the low impact of these factors in multivariate analyses has prevented their inclusion in the INRG risk classification schema (Cohn et al. 2009). The genetic aberrations of chromosomes 1p, 11q, and 17q are associated with poor outcome in neuroblastoma (Attiyeh et al. 2005; Bown et al. 1999; Caron et al. 1996; Cohn et al. 2009). Despite an increasing consideration in the clinical setting respect to the past is evident, these chromosome aberrations were only reported in one of the datasets (Figure 1). Nevertheless, both aberration 1p and 17q are statistically correlated with nMYC amplification: this correlation is the main reason why the INRG classification does not utilize them, and why under the proper circumstances nMYC amplification could be used as a proxy for these chromosomal aberrations (O'Neill et al. 2001).

Primary tumor sites and metastases. The location of the primary tumor and any metastatic sites dictates the symptomatology (Tolbert & Matthay 2018). Adrenal medulla is the most common site to develop a primary tumor, but tumor may arise also from the paraspinal or other sympathetic ganglia and can be present anywhere from the neck to the pelvis (Tolbert & Matthay 2018). Metastases are present at diagnosis in about 50% of patients, with the bone marrow, bone and regional lymph nodes being the most common sites (Tolbert & Matthay 2018). Two of out five datasets (Figure 1) reported the primary or metastasis sites. Neuroblastoma patients' outcome, indicating the patients' status alive or dead, and overall survival, that is the time interval between the last follow-up date and the date of diagnosis, are the primary endpoints of the whole clinical activity and biomarkers have been proposed in the literature to improve survival of patients (Cangelosi et al. 2020; Cangelosi et al. 2014, 2016, 2013; Ognibene et al. 2017). Two of the datasets have reported the outcome and the overall survival. The study by Kim et al. (2018) used the feature outcome but they refer to the onset of a relapse.

Relapse. Neuroblastoma patients' relapse indicates whether patients experienced a relapse. Relapse-free survival is the time interval between the date of first relapse and the date of diagnosis (Cangelosi et al. 2013). Currently, no curative salvage regimens for recurrent neuroblastoma are known, thus several studies have been reported in the literature to fill this gap (Basta et al. 2016). Only one of the five datasets' studies (Figure 1) reported features about relapse, relapse-free survival or relapse sites. Relapse-free survival was referred with progression free survival feature by Banelli et al. (2013). Methylation SFN, Methylation PCDHB cluster, HDCT2 regimen, HDCT1 regimen, and age at relapse are study-specific features and their role on neuroblastoma development or progression still remain to be validated.

Set of patients feature refers to a technical split of a cohort into two distinct subsets of patients for training and validate a computation model. One study reported this kind of feature.

Recap. Taken together, previous evidences suggest that:

- **(i)** The number and type of clinical features is heterogeneous across the five datasets' studies;
- **(ii)** The datasets' studies investigated distinct subsets of patients, which included high-risk patients with stage 4 tumor (Banelli et al. 2013), patients younger than 18 months, stage 4 and MYCN not amplified tumor (Kim et al. 2018), patients with MYCN amplified and low 11q tumor (Villamón et al. 2013), patients of all stages (Choi et al. 2019) and patients with relapsed or refractory disease after HDCT or auto-SCT (Ma et al. 2018);
- **(iii)** The features collected in previous studies are taken at different time points including diagnosis, treatment, patient's response, time of relapse, and follow-up);
- **(iv)** Sometimes different names are reported to refer to the same clinical parameter, and this is the case of *treatment response* (Villamón et al. 2013) and *tumor response* (Kim et al. 2018) or *time to first relapse* (Villamón et al. 2013) and *interval to relapse* (Choi et al. 2019);

(v) Once, the same feature name is reported to indicates different types of data as it is the case of *outcome* in the dataCK2018 (Kim et al. 2018) and *outcome* of dataBB2013 dataset (Banelli et al. 2013).

3 ANALYSIS

Neuroblastoma is the most common malignant tumors diagnosed in children under one year old, and is derived from the sympathetic nervous system. The clinical presentation of neuroblastoma patients is often vague, with many signs and symptoms being non-specific. Initial clinical presentation frequently includes weight loss, fever, and lethargy, with the most overt clinical signs and symptoms like a unilateral mass and opsoclonus myoclonos not showing up until much later in the disease progression (Keikhaei et al. 2012). Neuroblastoma is a tumor with clinical and prognostic heterogeneity; for some patients, the disease is an often aggressive and terminal disease, while other patients develop a benign disease which often has complete and spontaneous regression (Brodeur & Bagatell 2014). When diagnosed early as a localized disease, patients suffering from neuroblastoma frequently have high survival rates, approaching 90% in stage I and stage IVS of the disease. The survival rate, however, gets quite dismal in patients with stage III, with only 21% survival rate at 10 years (Bernstein et al. 1992).

Epidemiologically, neuroblastoma is a rare disease with an incidence of around 11 cases per million children (Spix et al. 2006). The heterogeneity of clinical presentation coupled with a low incidence rate make most clinical centers lack of data points to carry out meaningful clinical studies. Therefore, pooling data from multiple centers is critical for the development of new therapies and management guidelines. Share of patients' data across many medical centers have led to breakthrough studies in the past: Cohn and Pearson (Cohn et al. 2009) led a large consortium that recollected data from 8,800 patients suffering from neuroblastoma, and further analysis of this dataset led to the development of a new tumor staging system on the basis of surgical risk factors (Maris 2010). Risk stratification is pivotal in patients suffering from neuroblastoma, as it aids in choosing the most proper clinical management. Patients belonging to the low-risk strata are usually assigned to undergo local surgery resection; in contrast, patients in the high-risk strata are usually treated with high doses of chemotherapy, surgery, external beam therapy, and anti-GD2 immunotherapy (Maris 2010).

Even if stratification of neuroblastoma patients is pretty clear and concise at both end of the risk spectrum, however, there has not been a coherent consensus into how to accurately stratify those patients at an intermediate risk. Great efforts have been made to create a uniform risk stratification to share among research groups, the most important is the International Neuroblastoma Risk Group (INRG) classification system (Cohn et al. 2009). This risk stratification score utilizes age, histology, grade of tumor differentiation, MYCN, 11q aberration, and ploidy to assess the survival risk in pre-treatment patients.

In this research study, we found five open unrestricted datasets which have a diverse set of features that include clinical, genetic and laboratory data. Clinical data includes age and sex: age showing up in every data set and sex present in three out of five. Age at diagnosis is one of the most important factors at the moment of defining prognosis in neuroblastoma patients and is treated as a proxy for the underlying genetic and biologic features of the tumor (Cheung et al. 2012). Infants less than one year old usually suffer a far more a benign course than their older counterparts, with close to 80% survival rate, while adolescents suffer, for the most part, an aggressive and terminal disease with abysmal survival rates.

Clinical lab findings are missing in most data sets, with only dataCK2018 containing most of the lab findings like ferritin, LDH, NSE, and Urine VMA. This aspect can be concerning as these labs are for the most part are readily accessible even in underprivileged area, more so they are a key component in the work up of patients suffering from this oncologic disease. These markers are not routinely used in risk stratification neuroblastoma patients, as they are not as precise as the genetic markers (Sokol & Desai 2019); nevertheless, further inspection into the possible interactions with genetic and clinical values might shed a light into a more precise risk stratification.

It is crucial to further analyze how ferritin and LDH might aid in subgroup analysis for both prognosis and treatment, as it has been shown in the past that LDH aids on selecting the

appropriate therapeutics in other type of cancers. For example, initial levels of LHD can predict benefit of bevacizumab in colorectal cancer (Yin et al. 2014), and LDH response to first line treatment predicts survival in breast cancer (Pelizzari et al. 2019).

Clinical markers are insufficient for a precise risk stratification and subsequent treatment, and this is why the reason obtaining genetic data is crucial at the moment of managing patients with neuroblastoma. Analysis of cytogenetic profiles of neuroblastoma cells has revealed that whole chromosomal changes without any segmental alteration is associated with great outcomes even in older patients, or disseminated disease, meanwhile segmental chromosome imbalances are associated with worse prognosis and high risk of relapse, even in patients that presents whole chromosomal changes (Janoueix-Lerosey et al. 2009). MYCN amplification, a crucial genetic marker for risk stratification (Cao et al. 2017), is present in three out of five datasets. Other common genetic markers, such as ploidy and 11q abberations, which are used in the INGR risk stratification score, are severely lacking in the datasets, with 11q only being present in the dataCK2018 and ploidy missing completely.

Confirming the diagnosis of neuroblastoma requires a biopsy, which also brings important information for the prognosis. These tumors are classified depending on the amount of Schwannian stroma present in the tumor (Shimada et al. 2001) and, generally speaking, undifferentiated histology confers a worse prognosis, while highly differentiate histology confers a good prognosis. Histopathological information is present in the dataBB2013, dataEV2013 and dataYM2018 datasets, and can help researchers to further elucidate how histology of this tumor might interact with genomics and clinical markers.

Treatment in neuroblastoma patients has a binary trend due to its heterogeneity, the tendency in the past years has been to reduce therapy in patients belonging to the low-risk strata, while increasing it in those at the high-risk end. The focus in the past decade has been to use higher doses of chemotherapy and radiotherapy in patients with high risk neuroblastoma (Haghiri et al. 2021). Chemotherapy data is present in dataYBC2019, dataCK2018, and dataEV2013 datasets, and radiotherapy data is included in dataYM2018 and dataCk2018 cohorts, along with clinical, genomic and outcome data this could be used to support more individualized clinical decision making at the moment of picking the right treatment in neuroblastoma patients (Kent et al. 2018).

Limitations Integration of the datasets is possible due to presence of a subset of features that are reported for the majority of the datasets, such as age, stage, and MYCN status. These variables can be used for clinical and molecular characterization of the patients and their diseases.

However, the scarce overlap among the features across all datasets and the differences on the scale of feature values represent a limitation for these datasets, since this might reduce the comparability among the datasets and their usability in future studies of new biomedical markers for neuroblastoma. Compatibility across datasets is an important characteristic of the dataset and an aim of any standardization effort put in place by the scientific community. We believe that a larger set of common features would be of great benefit for the scientific community that will improve comparability across publicly available datasets and enhance the reusability of data for future studies on neuroblastoma.

The low number of patients included in a study is one of the most limiting factors in studies on rare diseases such as neuroblastoma. Three out of five datasets reported features for less than twenty neuroblastoma patients. The low number of patients represents an additional limitation of these datasets because it might reduce the robustness of the analysis be carried out on these datasets in future studies. A larger number of patients would be necessary to report robust conclusions and to achieve a sufficient trustworthy analysis able to support clinical decisions.

4 CONCLUSIONS

Neuroblastoma is a child cancer that affects thousands of newborns worldwide, and scientific research on the data derived from electronic health records can reveal new discoveries about this disease. EHR neuroblastoma datasets are available in the internet, but located on different websites or among supplementary information of published articles and therefore difficult

to find. We alleviate this issue by presenting here this survey where we describe each of the five public datasets on neuroblastoma EHRs currently available in the scientific literature, by reporting the quantitative characteristics and the clinical features of the five analyzed datasets, and highlighting which important variables were present in the datasets and which ones were absent (subsection S1.1).

In this survey, we also introduce our Neuroblastoma EHRs Open Data Repository, an online catalogue containing the five datasets which can be used by researchers and computers worldwide for any scientific analysis. Unlike the INRG Data Commons (Volchenboum et al. 2017) and the Pediatric Cancer Data Commons (Plana et al. 2021), our repository's data can be accessed openly without any specific permission or project proposal, by anyone in the world at any time.

We believe our survey and data repository can be useful resources that facilitate new scientific discoveries about neuroblastoma and that can lead to an improvement of the conditions of the patients worldwide.

In the future, we plan to write surveys and to develop data repositories derived from electronic health records of patients with other diseases, such as sepsis (Chicco & Jurman 2020) or amyotrophic lateral sclerosis (Kueffner et al. 2019).

S1 SUPPLEMENTARY INFORMATION

S1.1 ADDITIONAL CONSIDERATIONS

Initial work up of a recent patient diagnosed with neuroblastoma usually include a complete blood lab, with coagulogram, uric acid, electroyles, kidney function, liver function test, ferritin and LDH. LDH and Ferritin are both correlated with a poor prognosis when elevated (Hann et al. 1985; Quinn et al. 1980). These bloods tests are not included in the INRG classification system due to lack of specificity (Sokol & Desai 2019), nonetheless the employment of these markers might enhance the granularity of risk stratification while adding a negligent overhead to the management of patients; these lab tests are affordable and readily accessible even in underprivileged areas of the world. LDH can only be found in the dataCK2018 dataset, while ferritin is present in both dataBB2013 and dataCK2018.

Patients suspected of suffering of neuroblastoma should have a urine specimen recollected, since high levels of cathecolamis and their downstream metabolites are often elevated and support the diagnosis. Vanillylmandelic acid (VMA) and homovanillic acid (HVA) are both catechoalime metabolites that can be found in urine in up to 90% of children with neuroblastoma (Strenger et al. 2007). There are some information that these metabolites might aid in prognosis, especially when taking into account the DA (Dopamine)/VMA ratio, might help in biological grading (Strenger et al. 2007).

Neuron-specific enolase (NSE) a specific marker for neurons and peripheral neuroendocrine cells, have been found to be increased in advanced neuroblastoma and is correlated with poorer prognosis (Georgantzi et al. 2018). NSE is only available in the dataCK2018 dataset, and it might also help with sub-group analysis.

Various segmental chromosomal aberrations have been found to be associated with poor prognosis. Patients with loss of heterozygosity at 11q and 1p (Attiyeh et al. 2005) and gain at 17q (Lastowska et al. 1997) have been linked with poorer prognosis, only 11q is included in the INRG classification, mostly because 1p loss and 17q gain are statically associated with nMYC amplification (O'Neill et al. 2001). These genetic markers have been a key addition to risk stratification in the last decades, being nMYC amplification one of the strongest predictors for high-risk disease.

DATA ACCESSIBILITY STATEMENT

The five datasets described in this study are publicly available, under the CC BY 4.0 license (Creative Commons 2022), at the following web address:

https://davidechicco.github.io/neuroblastoma_EHRs_data or at

https://doi.org/10.5281/zenodo.6915403

Banelli, B, Bonassi, S, Casciano, I, Mazzocco, K, Di Vinci, A, Scaruffi, P, Brigati, C, Allemanni, G, Borzì, L, Tonini, GP and **Romani, M.** 2010. Outcome prediction and risk assessment by quantitative pyrosequencing methylation analysis of the SFN gene in advanced stage, high-risk, neuroblastic tumor patients. *International Journal of Cancer,* 126(3): 656–668. DOI: https://doi.org/10.1002/ijc.24768

Banelli, B, Di Vinci, A, Gelvi, I, Casciano, I, Allemanni, G, Bonassi, S and **Romani, M.** 2005a. DNA methylation in neuroblastic tumors. *Cancer Letters,* 228(1–2): 37–41. DOI: https://doi.org/10.1016/j.canlet.2005.02.049

Banelli, B, Gelvi, I, Di Vinci, A, Scaruffi, P, Casciano, I, Allemanni, G, Bonassi, S, Tonini, GP and **Romani, M.** 2005b. Distinct CpG methylation profiles characterize different clinical groups of neuroblastic tumors. *Oncogene*, 24(36): 5619–5628. DOI: https://doi.org/10.1038/sj.onc.1208722

Banelli, B, Merlo, DF, Allemanni, G, Forlani, A and **Romani, M.** 2013. Clinical potentials of methylator phenotype in stage 4 high-risk neuroblastoma: an open challenge. *PLOS One*, 8(5): e63253. DOI: https://doi.org/10.1371/journal.pone.0063253

Barco, S, Gennai, I, Reggiardo, G, Galleni, B, Barbagallo, L, Maffia, A, Viscardi, E, De Leonardis, F, Cecinati, V, Sorrentino, S, Garaventa, A, Conte, M and **Cangemi, G.** 2014. Urinary homovanillic and vanillylmandelic acid in the diagnosis of neuroblastoma: Report from the Italian Cooperative Group for Neuroblastoma. *Clinical Biochemistry*, 47(9): 848–852. DOI: https://doi.org/10.1016/j.clinbiochem.2014.04.015

Basta, NO, Halliday, GC, Makin, G, Birch, J, Feltbower, R, Bown, N, Elliott, M, Moreno, L, Barone, G, Pearson, AD, James, PW, Tweddle, DA and **McNally, RJ.** 2016. Factors associated with recurrence and survival length following relapse in patients with neuroblastoma. *British Journal of Cancer,* 115(9): 1048–1057. DOI: https://doi.org/10.1038/bjc.2016.302

Bernstein, ML, Leclerc, JM, Bunin, G, Brisson, L, Robison, L, Shuster, J, Byrne, T, Gregory, D, Hill, G and **Dougherty, G.** 1992. A population-based study of neuroblastoma incidence, survival, and mortality in North America. *Journal of Clinical Oncology*, 10(2): 323–329. DOI: https://doi.org/10.1200/JCO.1992.10.2.323

Bertagnolli, MM, Sartor, O, Chabner, BA, Rothenberg, ML, Khozin, S, Hugh-Jones, C, Reese, DM and **Murphy, MJ.** 2017. Advantages of a truly open-access data-sharing model. *New England Journal of Medicine,* 376(12): 1178–1181. DOI: https://doi.org/10.1056/NEJMsb1702054

Bown, N, Cotterill, S, Łastowska, M, O'Neill, S, Pearson, AD, Plantaz, D, Meddeb, M, Danglot, G, Brinkschmidt, C, Christiansen, H, Laureys, G, Nicholson, J, Bernheim, A, Betts, DR, Vandesompele, J, Van Roy, N and **Speleman, F.** 1999. Gain of chromosome arm 17q and adverse outcome in patients with neuroblastoma. *New England Journal of Medicine,* 340(25): 1954–1961. DOI: https://doi.org/10.1056/NEJM199906243402504

Brodeur, GM and **Bagatell, R.** 2014. Mechanisms of neuroblastoma regression. *Nature Reviews Clinical Oncology*, 11(12): 704–713. DOI: https://doi.org/10.1038/nrclinonc.2014.168

CAMDA. 2017. Neuroblastoma data integration challenge. http://camda2017.bioinf.jku.at/doku.php/contest_dataset#neuroblastoma_data_integration_challenge URL visited on 18th January 2022.

Cangelosi, D, Blengio, F, Versteeg, R, Eggert, A, Garaventa, A, Gambini, C, Conte, M, Eva, A, Muselli, M and **Varesio, L.** 2013. Logic learning machine creates explicit and stable rules stratifying neuroblastoma patients. *BMC Bioinformatics,* 14(7): 1–20. DOI: https://doi.org/10.1186/1471-2105-14-S7-S12

Cangelosi, D, Morini, M, Zanardi, N, Sementa, AR, Muselli, M, Conte, M, Garaventa, A, Pfeffer, U, Bosco, MC, Varesio, L and **Eva, A.** 2020. Hypoxia predicts poor prognosis in neuroblastoma patients and associates with biological mechanisms involved in telomerase activation and tumor microenvironment reprogramming. *Cancers*, 12(9): 2343. DOI: https://doi.org/10.3390/cancers12092343

Cangelosi, D, Muselli, M, Parodi, S, Blengio, F, Becherini, P, Versteeg, R, Conte, M and **Varesio, L.** 2014. Use of attribute driven incremental discretization and logic learning machine to build a prognostic classifier for neuroblastoma patients. *BMC Bioinformatics*, 15(5): 1–15. DOI: https://doi.org/10.1186/1471-2105-15-S5-S4

Cangelosi, D, Pelassa, S, Morini, M, Conte, M, Bosco, MC, Eva, A, Sementa, AR and **Varesio, L.** 2016. Artificial neural network classifier predicts neuroblastoma patients' outcome. *BMC Bioinformatics,* 17(12): 83–93. DOI: https://doi.org/10.1186/s12859-016-1194-3

Cangemi, G, Reggiardo, G, Barco, S, Barbagallo, L, Conte, M, D'Angelo, P, Bianchi, M, Favre, C, Galleni, B, Melioli, G, Haupt, R, Garaventa, A and **Corrias, MV.** 2012. Prognostic value of ferritin, neuron-specific enolase, lactate dehydrogenase, and urinary and plasmatic catecholamine metabolites in children with neuroblastoma. *OncoTargets and Therapy*, 5: 417. DOI: https://doi.org/10.2147/OTT.S36366

Cao, Y, Jin, Y, Yu, J, Wang, J, Yan, J and **Zhao, Q.** 2017. Research progress of neuroblastoma related gene variations. *Oncotarget,* 8(11): 18444. DOI: https://doi.org/10.18632/oncotarget.14408

Caron, H, van Sluis, P, de Kraker, J, Bökkerink, J, Egeler, M, Laureys, G, Slater, R, Westerveld, A, Voute, P and **Versteeg, R.** 1996. Allelic loss of chromosome 1p as a predictor of unfavorable outcome in patients with neuroblastoma. *New England Journal of Medicine*, 334(4): 225–230. DOI: https://doi.org/10.1056/NEJM199601253340404

Cheung, N-KV, Zhang, J, Lu, C, Parker, M, Bahrami, A, Tickoo, SK, Heguy, A, Pappo, AS, Federico, S, Dalton, J, Cheung, IY, Ding, L, Fulton, R, Wang, J, Chen, X, Becksfort, J, Wu, J, Billups, CA, Ellison, D, Mardis, ER, Wilson, RK, Downing, JR, Dyer, MA and **St Jude Children's Research Hospital Washington University Pediatric Cancer Genome Project.** 2012. Association of age at diagnosis and genetic mutations in patients with neuroblastoma. *Journal of the American Medical Association*, 307(10): 1062–1071. DOI: https://doi.org/10.1001/jama.2012.228

Chicco, D and **Jurman, G.** 2020. Survival prediction of patients with sepsis from age, sex, and septic episode number alone. *Scientific Reports*, 10(1): 1–12, 2020. DOI: https://doi.org/10.1038/s41598-020-73558-3

Children with Cancer UK. 2021. Neuroblastoma overview. https://www.childrenwithcancer.org.uk/childhood-cancer-info/cancer-types/neuroblastoma/ URL visited on 11th November 2021.

Choi, YB, Son, MH, Cho, HW, Ma, Y, Lee, JW, Kang, E-S, Yoo, KH, Her, JH, Lim, O, Jung, M, Hwang, YK, Sung, KW and **Koo, HH.** 2019. Safety and immune cell kinetics after donor natural killer cell infusion following haploidentical stem cell transplantation in children with recurrent neuroblastoma. *PLOS One*, 14(12): e0225998. DOI: https://doi.org/10.1371/journal.pone.0225998

Clark, K, Vendt, B, Smith, K, Freymann, J, Kirby, J, Koppel, P, Moore, S, Phillips, S, Maffitt, D, Pringle, M, Tarbox, L and **Prior, F.** 2013. The Cancer Imaging Archive (TCIA): Maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6): 1045–1057. DOI: https://doi.org/10.1007/s10278-013-9622-7

Cohn, SL, Pearson, AD, London, WB, Monclair, T, Ambros, PF, Brodeur, GM, Faldum, A, Hero, B, Iehara, T, Machin, D, Mosseri, V, Simon, T, Garaventa, A, Castel, V and **Matthay, KK.** 2009. The International Neuroblastoma Risk Group (INRG) classification system: An INRG task force report. *Journal of Clinical Oncology*, 27(2): 289. DOI: https://doi.org/10.1200/JCO.2008.16.6785

Colon, NC and **Chung, DH.** 2011. Neuroblastoma. *Advances in Pediatrics*, 58(1): 297–311. DOI: https://doi.org/10.1016/j.yapd.2011.03.011

Creative Commons. 2022. Attribution-noncommercial 4.0 international (cc by-nc 4.0). https://creativecommons.org/licenses/by-nc/4.0/ URL visited on 26th July 2022.

Federico, SM, Allewelt, H, Spunt, SL, Hudson, MM, Wu, J, Billups, CA, Jenkins, J, Santana, VM, Furman, WL and **McGregor, LM.** 2015. Subsequent malignant neoplasms in pediatric patients initially diagnosed with neuroblastoma. *Journal of Pediatric Hematology/Oncology*, 37(1): e6. DOI: https://doi.org/10.1097/MPH.0000000000000148

Ferraro, S, Braga, F, Luksch, R, Terenziani, M, Caruso, S and **Panteghini, M.** 2020. Measurement of serum neuron-specific enolase in neuroblastoma: Is there a clinical role? *Clinical Chemistry*, 66(5): 667–675. DOI: https://doi.org/10.1093/clinchem/hvaa073

Francescatto, M, Chierici, M, Rezvan Dezfooli, S, Zandoná, A, Jurman, G and **Furlanello, C.** 2018. Multi-omics integration for neuroblastoma clinical endpoint prediction. *Biology Direct*, 13(1): 1–12. DOI: https://doi.org/10.1186/s13062-018-0207-8

Georgantzi, K, Sköldenberg, EG, Stridsberg, M, Kogner, P, Jakobson, A, Janson, ET and **Christofferson, RH.** 2018. Chromogranin A and neuron-specific enolase in neuroblastoma: correlation to stage and prognostic factors. *Pediatric Hematology and Oncology*, 35(2): 156–165. DOI: https://doi.org/10.1080/08880018.2018.1464087

Google. 2021. Google Scholar. https://scholar.google.com URL visited on 1st November 2011.

Google. 2022. Google Dataset Search. https://datasetsearch.research.google.com/ URL visited on 4th April 2022.

Haghiri, S, Fayech, C, Mansouri, I, Dufour, C, Pasqualini, C, Bolle, S, Rivollet, S, Dumas, A, Boumaraf, A, Belhout, A, Journy, N, Souchard, V, Vu-Bezin, G, Veres, C, Haddy, N, De Vathaire, F, Valteau-Couanet, D and **Fresneau, B.** 2021. Long-term follow-up of high-risk neuroblastoma survivors treated with high-dose chemotherapy and stem cell transplantation rescue. *Bone Marrow Transplantation*, 56: 1984–1997. DOI: https://doi.org/10.1038/s41409-021-01258-1

Hann, H-WL, Evans, AE, Siegel, SE, Wong, KY, Sather, H, Dalton, A, Hammond, D and **Seeger, RC.** 1985. Prognostic importance of serum ferritin in patients with Stages III and IV neuroblastoma: The Children's Cancer Study Group experience. *Cancer Research*, 45(6): 2843–2848.

International Neuroblastoma Risk Group. 2017. INRG Data Commons. https://inrgdb.org/ URL visited on 2nd March 2022.

Janoueix-Lerosey, I, Schleiermacher, G, Michels, E, Mosseri, V, Ribeiro, A, Lequin, D, Vermeulen, J, Couturier, J, Peuchmaur, M, Valent, A, Plantaz, D, Rubie, H, Valteau-Couanet, D, Thomas, C, Combaret, V, Rousseau, R, Eggert, A, Michon, J, Speleman, F and **Delattre, O.** 2009. Overall genomic pattern is a predictor of outcome in neuroblastoma. *Journal of Clinical Oncology*, 27(7): 1026–1033. DOI: https://doi.org/10.1200/JCO.2008.16.0630

Kaggle. 2022. Kaggle.com – Find open datasets. https://www.kaggle.com/datasets URL visited on 27th March 2022.

Keikhaei, B, Pedram, M, Popak, B, Heidari, M, Hadadi, N and **Samadi, B.** 2012. Signs and symptoms of neuroblastoma. *Journal of Medicine and Medical Science*, 3(4): 243–246.

Kent, DM, Steyerberg, E and **van Klaveren, D.** 2018. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. *BMJ*, 363. DOI: https://doi.org/10.1136/bmj.k4245

Khan, SM, Liu, X, Nath, S, Korot, E, Faes, L, Wagner, SK, Keane, PA, Sebire, NJ, Burton, MJ and **Denniston, AK.** 2021. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *The Lancet Digital Health*, 3(1): e51–e66. DOI: https://doi.org/10.1016/S2589-7500(20)30240-5

Kim, C, Choi, YB, Lee, JW, Yoo, KH, Sung, KW and **Koo, HH.** 2018. Excellent treatment outcomes in children younger than 18 months with stage 4 MYCN nonamplified neuroblastoma. *Korean Journal of Pediatrics*, 61(2): 53. DOI: https://doi.org/10.3345/kjp.2018.61.2.53

Kueffner, R, Zach, N, Bronfeld, M, Norel, R, Atassi, N, Balagurusamy, V, Di Camillo, B, Chio, A, Cudkowicz, M, Dillenberger, D, Garcia-Garcia, J, Hardiman, O, Hoff, B, Knight, J, Leitner, ML, Li, G, Mangravite, L, Norman, T, Wang, L, the ALS Stratification Consortium, Xiao, J, Fang, W-C, Peng, J, Yang, C, Chang, H-J and **Stolovitzky, G.** 2019. Stratification of amyotrophic lateral sclerosis patients: A crowdsourcing approach. *Scientific Reports*, 9(1): 690. DOI: https://doi.org/10.1038/s41598-018-36873-4

Lastowska, M, Cotterill, S, Pearson, A, Roberts, P, McGuckin, A, Lewis, I and **Bown, N.** 1997. Gain of chromosome arm 17q predicts unfavourable outcome in neuroblastoma patients. UK Children's Cancer Study Group and the UK Cancer Cytogenetics Group. *European Journal of Cancer*, 33(10): 1627–1633. DOI: https://doi.org/10.1016/S0959-8049(97)00282-7

Ma, Y, Zheng, J, Feng, J, Chen, L, Dong, K and **Xiao, X.** 2018. Neuroblastomas in Eastern China: A retrospective series study of 275 cases in a regional center. *PeerJ*, 6: e5665. DOI: https://doi.org/10.7717/peerj.5665

Maris, JM. 2010. Recent advances in neuroblastoma. *New England Journal of Medicine*, 362(23): 2202–2211. DOI: https://doi.org/10.1056/NEJMra0804577

Matsumura, T, Terada, J, Kinoshita, T, Sakurai, Y, Yahaba, M, Tsushima, K, Sakao, S, Nagashima, K, Ozaki, T, Kobayashi, Y, Hiwasa, T and **Tatsumi, K.** 2018. Circulating autoantibodies against neuroblastoma suppressor of tumorigenicity 1 (NBL1): A potential biomarker for coronary artery disease in patients with obstructive sleep apnea. *PLOS One*, 13(3): e0195015. DOI: https://doi.org/10.1371/journal.pone.0195015

Melaiu, O, Chierici, M, Lucarini, V, Jurman, G, Conti, LA, De Vito, R, Boldrini, R, Cifaldi, L, Castellano, A, Furlanello, C, Barnaba, V, Locatelli, F and **Fruci, D.** 2020. Cellular and gene signatures of tumor-infiltrating dendritic cells and natural-killer cells predict prognosis of neuroblastoma. *Nature Communications*, 11(1): 1–15. DOI: https://doi.org/10.1038/s41467-020-19781-y

Moroz, V, Machin, D, Hero, B, Ladenstein, R, Berthold, F, Kao, P, Obeng, Y, Pearson, AD, Cohn, SL and **London, WB.** 2020. The prognostic strength of serum LDH and serum ferritin in children with neuroblastoma: A report from the International Neuroblastoma Risk Group (INRG) project. *Pediatric Blood & Cancer*, 67(8): e28359. DOI: https://doi.org/10.1002/pbc.28359

National Health Institutes (NIH), Genetic and Rare Diseases Information Center (GARD). 2021. Chromosome 11q deletion. https://rarediseases.info.nih.gov/diseases/1735/chromosome-11q-deletion URL visited on 5th November 2021.

Noguera, R 7th November 2021 Personal communication (email).

Ognibene, M, Cangelosi, D, Morini, M, Segalerba, D, Bosco, MC, Sementa, AR, Eva, A and **Varesio, L.** 2017. Immunohistochemical analysis of PDK1, PHD3 and HIF-1*á* expression defines the hypoxic status of neuroblastoma tumors. *PLOS One*, 12(11): e0187206. DOI: https://doi.org/10.1371/journal.pone.0187206

O'Neill, S, Ekstrom, L, Lastowska, M, Roberts, P, Brodeur, GM, Kees, UR, Schwab, M and **Bown, N.** 2001. MYCN amplification and 17q in neuroblastoma: evidence for structural association. *Genes, Chromosomes and Cancer*, 30(1): 87–90. DOI: https://doi.org/10.1002/1098-2264(2000)9999:9999<::AID-GCC1055>3.0.CO;2-J

Park, JR, Bagatell, R, Cohn, SL, Pearson, AD, Villablanca, JG, Berthold, F, Burchill, S, Boubaker, A, McHugh, K, Nuchtern, JG, London, WB, Seibel, NL, Lindwasser, OW, Maris, JM, Brock, P, Schleiermacher, G, Ladenstein, R, Matthay, KK and **Valteau-Couanet, D.** 2017. Revisions to the international neuroblastoma response criteria: a consensus statement from the national cancer institute clinical trials planning meeting. *Journal of Clinical Oncology*, 35(22): 2580. DOI: https://doi.org/10.1200/JCO.2016.72.0177

Pautasso, M. 2013. Ten simple rules for writing a literature review. *PLOS Computational Biology*, 9(7): e1003149. DOI: https://doi.org/10.1371/journal.pcbi.1003149

Pelizzari, G, Basile, D, Zago, S, Lisanti, C, Bartoletti, M, Bortot, L, Vitale, MG, Fanotto, V, Barban, S, Cinausero, M, Bonotto, M, Gerratana, L, Mansutti, M, Curcio, F, Fasola, G, Minisini, AM and **Puglisi, F.** 2019. Lactate dehydrogenase (LDH) response to first-line treatment predicts survival in metastatic breast cancer: First clues for a cost-effective and dynamic biomarker. *Cancers*, 11(9): 1243. DOI: https://doi.org/10.3390/cancers11091243

Plana, A, Furner, B, Palese, M, Dussault, N, Birz, S, Graglia, L, Kush, M, Nicholson, J, Hecker-Nolting, S, Gaspar, N, Rasche, M, Bisogno, G, Reinhardt, D, Zwaan, CM, Koscielniak, E, Frazier, AL, Janeway, K, Hawkins, DS, Kolb, EA, Cohn, SL, Pearson, ADJ and **Volchenboum, SL.** 2021. Volchenboum. Pediatric cancer data commons: federating and democratizing data for childhood cancer research. *JCO Clinical Cancer Informatics*, 5: 1034–1043. DOI: https://doi.org/10.1200/CCI.21.00075

Qi, Y and **Zhan, J.** 2021. Roles of surgery in the treatment of patients with high-risk neuroblastoma in the children oncology group study: A systematic review and meta-analysis. *Frontiers in Pediatrics*, 9: 1–9. DOI: https://doi.org/10.3389/fped.2021.706800

Quinn, JJ, Altman, AJ and **Frantz, CN.** 1980. Serum lactic dehydrogenase, an indicator of tumor activity in neuroblastoma. *Journal of Pediatrics*, 97(1): 89–91. DOI: https://doi.org/10.1016/S0022-3476(80)80139-9

Re3data. 2022. Registry of research data repositories. https://www.re3data.org/ URL visited on 24th June 2022.

Romani, M. 8th November 2021. Personal communication (email).

Rosenbaum, DG, Abramson, SJ, DeLappe, E, Teruya-Feldstein, J, La Quaglia, MP, Fox, JJ and **Price, AP.** 2013. Pancreatic involvement in neuroblastoma with radiologic-pathologic correlation: A single-institution experience. *American Journal of Roentgenology*, 201(1): W141–W146. DOI: https://doi.org/10.2214/AJR.12.9618

Shimada, H, Umehara, S, Monobe, Y, Hachitanda, Y, Nakagawa, A, Goto, S, Gerbing, RB, Stram, DO, Lukens, JN and **Matthay, KK.** 2001. International neuroblastoma pathology classification for prognostic evaluation of patients with peripheral neuroblastic tumors: A report from the children's cancer group. *Cancer*, 92(9): 2451–2461. DOI: https://doi.org/10.1002/1097-0142(20011101)92:9<2451::AID-CNCR1595>3.0.CO;2-S

Smith, SJ, Diehl, NN, Smith, BD and **Mohney, BG.** 2010. Urine catecholamine levels as diagnostic markers for neuroblastoma in a defined population: Implications for ophthalmic practice. *Eye*, 24(12): 1792–1796. DOI: https://doi.org/10.1038/eye.2010.125

Software Carpentry. 2022. Reading and writing CSV files. https://swcarpentry.github.io/r-novice-inflammation/11-supp-read-write-csv/ URL visited on 27th March 2022.

Sokol, E and **Desai, AV.** 2019. The evolution of risk classification for neuroblastoma. *Children*, 6(2): 27. DOI: https://doi.org/10.3390/children6020027

Spix, C, Pastore, G, Sankila, R, Stiller, CA and **Steliarova-Foucher, E.** 2006. Neuroblastoma incidence and survival in European children (1978–1997): Report from the Automated Childhood Cancer Information System project. *European Journal of Cancer*, 42(13): 2081–2091. DOI: https://doi.org/10.1016/j.ejca.2006.05.008

Stall, S, Yarmey, L, Cutcher-Gershenfeld, J, Hanson, B, Lehnert, K, Nosek, B, Parsons, M, Robinson, E and **Wyborn, L.** 2019. Make scientific data FAIR. *Nature*, 570(7759): 27–29. DOI: https://doi.org/10.1038/d41586-019-01720-7

Strenger, V, Kerbl, R, Dornbusch, HJ, Ladenstein, R, Ambros, PF, Ambros, IM and **Urban, C.** 2007. Diagnostic and prognostic impact of urinary catecholamines in neuroblastoma patients. *Pediatric Blood & Cancer*, 48(5): 504–509. DOI: https://doi.org/10.1002/pbc.20888

The Document Foundation. 2022. LibreOffice Calc. https://www.libreoffice.org/discover/calc/ URL visited on 18th January 2022.

Tolbert, VP and **Matthay, KK.** 2018. Neuroblastoma: Clinical and biological approach to risk stratification and treatment. *Cell and Tissue Research*, 372(2): 195–209. DOI: https://doi.org/10.1007/s00441-018-2821-2

Trahair, TN, Vowels, MR, Johnston, K, Cohn, RJ, Russell, S, Neville, KA, Carroll, S and **Marshall, GM.** 2007. Long-term outcomes in children with high-risk neuroblastoma treated with autologous stem cell transplantation. *Bone Marrow Transplantation*, 40(8): 741–746. DOI: https://doi.org/10.1038/sj.bmt.1705809

University of California Irvine. 1987. Machine Learning Repository. https://archive.ics.uci.edu/ml URL visited on 8th November 2021.

US National Center for Biotechnology Information. 2021. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo URL visited on 8th November 2021.

Villamón, E, Berbegall, AP, Piqueras, M, Tadeo, I, Castel, V, Djos, A, Martinsson, T, Navarro, S and **Noguera, R.** 2013. Genetic instability and intratumoral heterogeneity in neuroblastoma with MYCN amplification plus 11q deletion. *PLOS One*, 8(1): e53740. DOI: https://doi.org/10.1371/journal.pone.0053740

Volchenboum, S, Cohen, E, Furner, B and **the Pediatric Cancer Data Commons Team.** 2021. Pediatric Cancer Data Commons. https://commons.cri.uchicago.edu/ URL visited on 27th March 2022.

Volchenboum, SL, Cox, SM, Heath, A, Resnick, A, Cohn, SL and **Grossman, R.** 2017. Data commons to support pediatric cancer research. *American Society of Clinical Oncology Educational Book*, 37: 746–752, 2017. DOI: https://doi.org/10.1200/EDBK_175029

Wickham, H. 2016. Programming with ggplot2. In *ggplot2*, pages 241–253. Springer. DOI: https://doi.org/10.1007/978-3-319-24277-4_12

Wilkinson, M, Dumontier, M, Aalbersberg, I, Appleton, G, Axton, M, Baak, A, Blomberg, N, Boiten, J, da Silva Santos, L, Bourne, P, Bouwman, J, Brookes, A, Clark, T, Crosas, M, Dillo, I, Dumon, O, Edmunds, S, Evelo, C, Finkers, R, Gonzalez-Beltran, A, Gray, A, Groth, P, Goble, C, Grethe, J, Heringa, J, 't Hoen, P, Hooft, R, Kuhn, T, Kok, R, Kok, J, Lusher, S, Martone, M, Mons, A, Packer, A, Persson, B, Rocca-Serra, P, Roos, M, van Schaik, R, Sansone, S, Schultes, E, Sengstag, T, Slater, T, Strawn, G, Swertz, M, Thompson, M, Van Der Lei, J, Van Mulligen, E, Velterop, J, Waagmeester, A, Wittenburg, P, Wolstencroft, K, Zhao, J and **Mons, B.** 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3: 160018. DOI: https://doi.org/10.1038/sdata.2016.18

Williams, LA, Richardson, M, Marcotte, EL, Poynter, JN and **Spector, LG.** 2019. Sex ratio among childhood cancers by single year of age. *Pediatric Blood & Cancer*, 66(6): e27620. DOI: https://doi.org/10.1002/pbc.27620

Yin, C, Jiang, C, Liao, F, Rong, Y, Cai, X, Guo, G, Qiu, H, Chen, X, Zhang, B, He, W and **Xia, L.** 2014. Initial LDH level can predict the survival benefit from bevacizumab in the first-line setting in Chinese patients with metastatic colorectal cancer. *OncoTargets and Therapy*, 7: 1415. DOI: https://doi.org/10.2147/OTT.S64559

TO CITE THIS ARTICLE:
Chicco, D, Cerono, G and Cangelosi, D. 2022. A Survey on Publicly Available Open Datasets Derived From Electronic Health Records (EHRs) of Patients with Neuroblastoma. *Data Science Journal*, 21: 17, pp. 1–18. DOI: https://doi.org/10.5334/dsj-2022-017

Submitted: 03 May 2022
Accepted: 19 August 2022
Published: 04 October 2022

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.