The PDF Data Extractor (PDE) Pre-screening Tool Reduced the Manual Review Burden for Systematic Literature Reviews by Over 35% Through Automated High-Throughput Assessment of Full-Text Articles

Erik Stricker[1§*] **and Michael E. Scheurer**[1§*]

[1]Baylor College of Medicine

***For correspondence:**
stricker@bcm.edu (ES);
scheurer@bcm.edu (MES)

Present address: [§]Department of Pediatrics, Department of Molecular Virology and Microbiology
Baylor College of Medicine, Houston, TX 77030, USA

Abstract Literature reviews are generally time-consuming and rely heavily on accurate representation of the data in the title and abstract of articles. Often minor results and details are lost in a systematic screen, which is becoming even more frequent with the rapidly rising numbers of daily published scientific articles. We developed the PDF Data Extractor (PDE) R package to aid scientists at any stage in literature reviews while offering a user-friendly interface. The tool permits the user to categorize large numbers of full-text articles in PDF format, export containing tables to Excel sheets (pdf2table), and extract relevant data using a simple user interface, requiring no bioinformatics skills. Specific features of the literature analysis comprise the adaptability of analysis parameters including the use of regular expressions, machine learning-powered detection of abbreviations of search words in articles, and the export of document meta-data. We exemplify how the PDE R package can be utilized as a pre-screening tool allowing automated categorization of full-text articles by relevance, thereby reducing the literature to be evaluated (in our example by 35% with a sensitivity of 100% at standard parameters). The PDE R package is available from the Comprehensive R Archive Network at https://CRAN.R-project.org/package=PDE and as web tool with limited capacity at https://erikstricker.shinyapps.io/PDE_analyzer/.

Introduction

In an age of exponentially increasing numbers of published scientific articles, it is surprising that most systematic literature reviews are still conducted manually evaluating each article individually. Systematic literature reviews aim to find and collect relevant information concerning a specific research question and are an essential step in virtually every area of research, e.g., for the preparation of review articles, project proposals, and experimental designs. While machine learning tools are available for literature searches and screens (***Marshall and Wallace, 2019***), they: 1) require a large number of manually evaluated articles for the training of the tool, 2) are often restricted to

filtering articles by study design or choosing topics from a limited set of terms, and 3) are generally limited to the evaluation of article titles and abstracts.

Furthermore, extraction of tabularized data from PDF files poses an additional challenge. Existing R packages such as pdftools (***Ooms, 2019***) or tabulizer (***Leeper, 2018***) require programming and data conversion skills, deterring many users. In addition, table extraction with these packages requires the tables to either be the only element on a page or be manually selected for each instance. While many online tools allow the fast and relatively accurate conversion of PDF files into Microsoft Excel files, they require a paid subscription for the processing of more than a few files (e.g., https://smallpdf.com/, https://www.adobe.com/acrobat/online/pdf-to-excel.html, https://pdftables.com/) or are limited in input file numbers and, therefore, do not allow high-throughput table extraction (e.g., https://pdftoxls.com/, https://docs.zone/pdf-to-excel, https://www.pdftoexcel.com/).

Addressing these limitations, we developed the openly available, free-to-use PDE R package with user-friendly interface. The PDF data extractor (PDE) R package easily extracts information and tables from full-text articles in Portable Document Format (PDF) format based on user-defined keywords and does not require a training set. The output tables are saved in universal comma-separated values (.csv) or tab-separated values (.tsv) files easily opened in Microsoft Excel. Additional features include the adaptability of analysis parameters including the use of regular expressions, machine learning powered detection of abbreviations of search words in articles, and the export of document meta-data. The PDE R package is available from the Comprehensive R Archive Network at https://CRAN.R-project.org/package=PDE).

Results

Implementation of PDE R Package as Pre-screening Tool Significantly Reduced Manual Screening Requirements

In their Cochrane systematic review, Kraal *et al.* searched the Cochrane Central Register of Controlled Trials (CENTRAL; the Cochrane Library 2016, Issue 3), PubMed (1945 to 25 April 2016) and Embase (Ovid) (1980 to 25 April 2016) for articles on 131I-MIBG and HR NBL (***Kraal et al., 2015***). All article titles and abstracts were manually evaluated by the group and eligible articles were assessed in the full text. To simplify comparisons between the traditional and a PDE-facilitated approach, we concentrated our evaluations on the articles found in the PubMed database which were 2,291 out of the 3,366 articles screened by Kraal *et al.* (see Fig. 1). We were able to obtain 1262 full-text articles in PDF format and used them for pre-screening facilitated by the PDE R package. The analysis took approximately 1.5 hours with an average of 4 seconds per article and yielded 785 articles containing the filter/search words detected at least 20-times and 24 articles without containing text that could be processed, i.e., secured articles or scans of full articles. Accordingly, over one third of the articles (35.9%, 454/1262) were excluded without the requirement of a manual review. The 762 articles not excluded encompassed 38/39 (97.4%) of the PubMed-derived articles that ***Kraal et al.*** (***2017***) assessed for eligibility. The 39 evaluated articles contained on average, 88 detected instances of filter words, with a range of 11 to 253, and an average of 56 sentences with search words extracted, ranging from 14 to 135. The two articles by ***de Kraker et al.*** (***2008***), and ***Kraal et al.*** (***2015***), finally included in the Cochrane systematic review displayed the filter words 72 and 89 times, with 50 and 31 sentences extracted, respectively.

The PDE R Package Displayed High Sensitivity in the Categorization of Full-Text Articles

An evaluation of specificity and sensitivity of the PDE R package was challenging since we observed that with higher thresholds for filter words, the selectivity was increased, whereas the number of articles assessed for eligibility in this pool were reduced although both included articles were still detected (Fig. 2). For example, a filter word threshold of 50 resulted in 376 articles for manual screening with and overlap of 19/39 full-text articles assessed by ***Kraal et al.*** (***2017***) and both in-

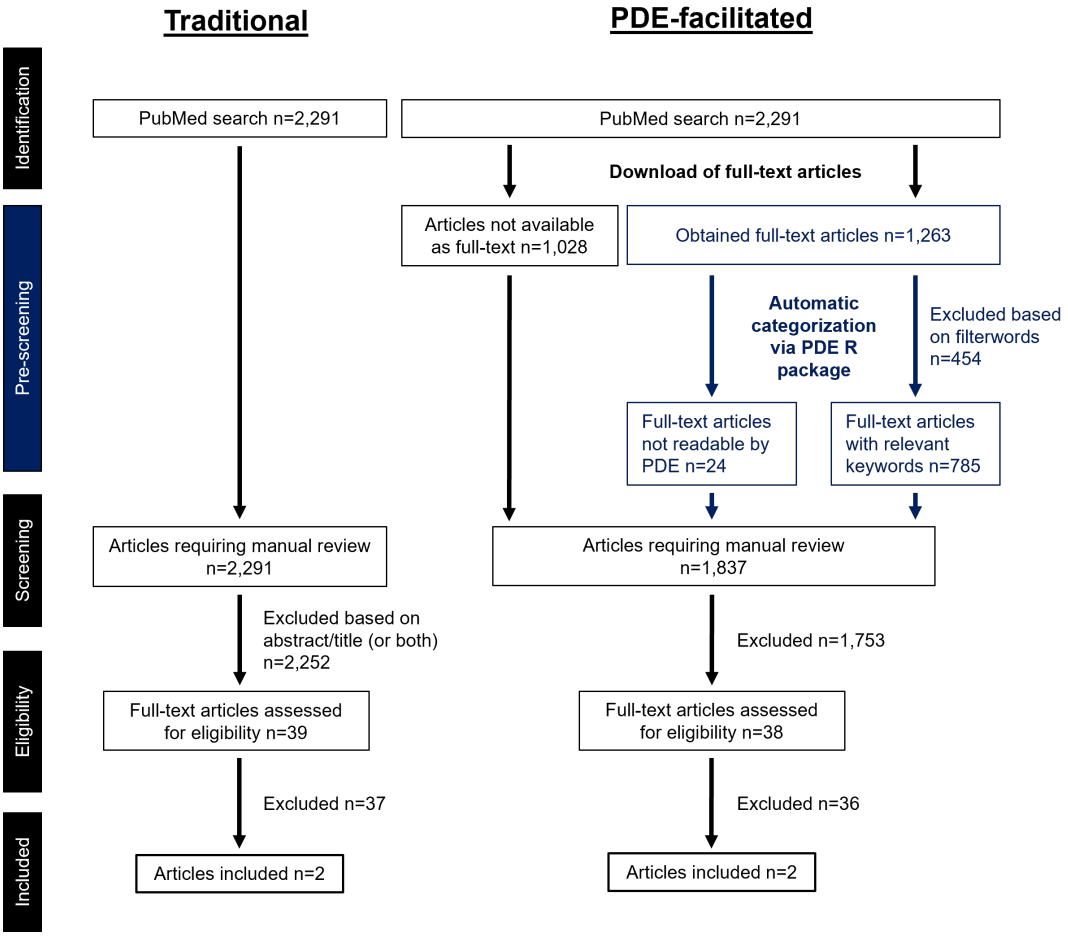


Figure 1. Modified PRISMA diagram comparison between the traditional approach used by ***Kraal et al.*** (***2017***) and the methodology using the PDE R package as prescreening tool (blue).

cluded articles still being detected. Overall, at standard parameters, we can report a specificity of 64% and sensitivity of 97-100% depending on whether full-text articles assessed for eligibility by ***Kraal et al.*** (***2017***) or articles included in the review are considered true positives. In summary, the PDE R package notably reduced the number of articles required to be manually reviewed with high accuracy, provided additional preprocessed content for the assessment of articles, i.e., extracted sentences with search words, and evaluated articles in a reproducible manner easily applied to additional literature.

Interestingly, the article by ***Mastrangelo*** (***1987***) was further evaluated by ***Kraal et al.*** (***2017***) but then excluded from the review due to a lack of primary data. The PDE R package, on the other hand, excluded the article during initial analysis based on the number of filter word number (fw=11 was below the threshold of 20) (see Supplemental Table S1). In contrast, the clinical study by ***Garaventa et al.*** (***1999***) was excluded from the Cochrane Review even though the abstract and title indicated 30 patients with stage 4 neuroblastoma while the PDE R package detected 127 filter words present. Furthermore, review of the 63 sentences extracted by the PDE R package highlighted observed adverse effects such as acute myeloid leukemia, thrombocytopenia, severe interstitial pneumonia, and reverse thyroid reserve (Supplemental Table S2) (***Garaventa et al., 1999***). Only when consulting the clinical table in ***Garaventa et al.*** (***1999***) it becomes evident that only four patients with stage 4 neuroblastoma displayed MYCN amplification with >8 copies thus not reaching the inclusion criteria of at least 10 patients with HR NBL in the study. Accordingly, it cannot be excluded that other studies were falsely excluded while the PDE R package would have prevented such a mistake.

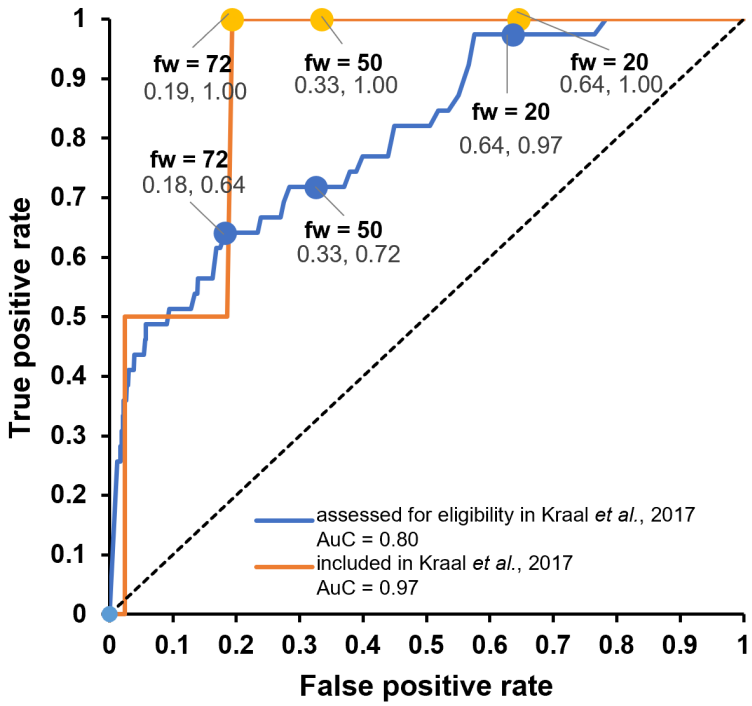
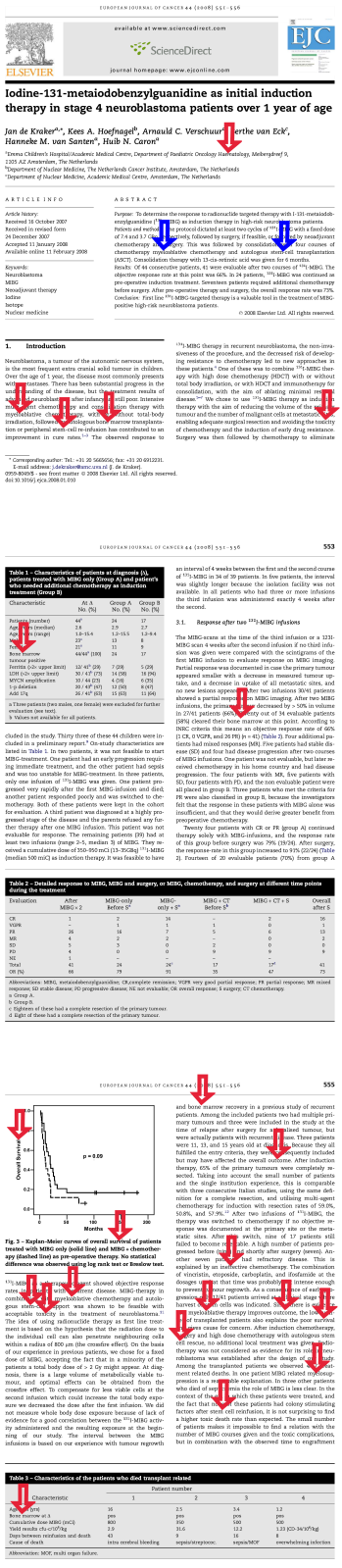


Figure 2. Receiver operating characteristic (ROC) curves of the article detection by the PDE_analyzer. ROC curves for the detection of the articles assessed for eligibility in ***Kraal et al.*** (***2017***) (blue) and the articles included in the review by ***Kraal et al.*** (***2017***) (orange) were compared to the reference line (dashed line). Variable parameter was the filter word threshold (fw) with different thresholds indicated by solid circles. The area under the curve (AuC) for each evaluation is indicated in the legend.

Direct comparison of abstract- and title-only review and the PDE R package showed that 97.2% (70/72) of the search words were located within the body article, only two search words in the abstract and none in the title (Fig. 3). The distribution of search words was even throughout the article by (***Mastrangelo, 1987***) with 4/5 figures/tables including search words. While 59 search words were located in the body of the article, 11 search words were in the references. The search term "surviv" was most common with 10 incidences (8 survival, 1 surviving, 1 survivors) although absent in title and abstract.

The PDE R Package Allows High-Throughput Table Extraction from PDF Files

Tables of a PDF document can be exported into a Microsoft Excel readable file format with and without the use of search words as exemplified in Figure 4. The PDE_analyzer detects the beginning of a table based on the standard annotation of tables in scientific literature, i.e., "Table" [Table index] [Table heading], or user-defined table headings, allowing the exclusion of non-table content from the PDF file. Tables that carry their heading below the actual table are not detected correctly due to the nature of the detection algorithm. To detect the content of each table cell and its location, an HTML copy of the PDF file is generated of which x and y coordinates for each text element can be extracted. For the assessment of the table structure, content starting with the fifth element in the table are compared in their coordinates. For this reason, tables with less than 5 elements cannot be processed by the PDE R package. After extraction of the table itself, the table legend is identified using font size differences in the document. The final tables are saved either as a comma- or tab-separated values file, i.e., *.csv or *.tsv.



EUROPEAN JOURNAL OF CANCER 44 (2008) 551–556

available at www.sciencedirect.com
ScienceDirect
journal homepage: www.ejconline.com

Iodine-131-metaiodobenzylguanidine as initial induction therapy in stage 4 neuroblastoma patients over 1 year of age

Jan de Kraker[a,*], Kees A. Hoefnagel[b], Arnauld C. Verschuur[c], Berthe van Eck[c], Hanneke M. van Santen[a], Huib N. Caron[a]

[1]Emma Children's Hospital/Academic Medical Centre, Department of Paediatric Oncology Haematology, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands
[2]Department of Nuclear Medicine, The Netherlands Cancer Institute, Amsterdam, The Netherlands
[3]Department of Nuclear Medicine, Academic Medical Centre, Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 16 October 2007
Received in revised form
24 December 2007
Accepted 11 January 2008
Available online 11 February 2008

Keywords:
Neuroblastoma
MIBG
Neoadjuvant therapy
Iodine
Isotope
Nuclear medicine

ABSTRACT

Purpose: To determine the response to radioactide targeted therapy with I-131-metaiodobenzylguanidine (I-131-MIBG) as induction therapy in high-risk neuroblastoma patients.
Patients and methods: A protocol dictated at least two cycles of I-131-MIBG with a fixed dose of 7.4 and 17.2 GBq, followed by surgery, if feasible, and (chemotherapy neoadjuvant chemotherapy and surgery. This was followed by consolidation with four courses of chemotherapy myeloablative chemotherapy and autologous stem cell transplantation (ASCT). Consolidation therapy with 13-cis-retinoic acid was given for 6 months.
Results: Of 40 consecutive patients, 41 were evaluable after two courses of ^{131}I-MIBG. The objective response rate at this point was 66%. In 24 patients, ^{131}I-MIBG was continued as pre-operative induction treatment. Seventeen patients required additional chemotherapy before surgery. After pre-operative therapy and surgery, the overall response rate was 73%.
Conclusion: First line ^{131}I-MIBG therapy is a valuable tool in the treatment of MIBG-positive high-risk neuroblastoma patients.

© 2008 Elsevier Ltd. All rights reserved.

1. Introduction

Neuroblastoma, a tumour of the autonomic nervous system, is the most frequent extra cranial solid tumour in children. Over the age of 1 year, the disease most commonly presents with metastases. There has been substantial progress in the understanding of the disease, but treatment results of high-risk neuroblastoma after infancy still poor. Intensive multimodal chemotherapy and consolidation therapy with myeloablative chemotherapy, with or without total-body irradiation, followed by autologous bone marrow transplantation or peripheral stem-cell re-infusion has contributed to an improvement in cure rates.[1–3] The observed response to ^{131}I-MIBG therapy in recurrent neuroblastoma, the non-invasiveness of the procedure, and the decreased risk of developing resistance to chemotherapy led to new approaches in these patients.[4] One of these was to combine ^{131}I-MIBG therapy with high dose chemotherapy (HDCT) with or without total body irradiation, or with HDCT and immunotherapy for consolidation, with the aim of debiting minimal residual disease.[5–7] We chose to use ^{131}I-MIBG therapy as induction therapy with the aim of reducing the volume of the primary tumour and the number of malignant cells at metastatic sites, enabling adequate surgical resection and avoiding the toxicity of chemotherapy and the induction of early drug resistance. Surgery was then followed by chemotherapy to eliminate residual disease. We present an analysis of 44 consecutively registered patients to investigate the feasibility and response to MIBG when used as induction treatment.

2. Patients and methods

Patients between 1 and 18 years of age at diagnosis who had high-risk, MIBG positive neuroblastoma were eligible for the study. The parents were extensively informed about the nature and precise elements of the treatment and the differences with other strategies and consent was obtained prior to inclusion. Only chemotherapy was an exclusion criterion, but initial surgery as the only treatment modality was not. Initial staging evaluation included computed tomography imaging of the primary tumour, ^{123}I-MIBG diagnostic scintigraphy, bone marrow aspirates, and biopsies of both sides of the iliac crest. The two most frequent genetic alterations, MYCN amplification and deletion of the chromosomal 1p region, were analysed. Lactate dehydrogenase (LDH) and ferritin were documented at diagnosis and later at specific time points. All patients were staged according to the revised International Neuroblastoma Staging System (INSS) and assessment of response was according to the International Neuroblastoma Response Criteria (INRC).[8] Survival distributions were calculated according the Kaplan–Meier method and differences in survival between groups were compared using Log rank and Breslow tests.

2.1. Study design

The first objective of the study was to document response after two infusions of MIBG as a single agent given after diagnosis in high-risk patients. The second objective was to find the rate of resectability of >50% of the primary tumour together with a significant reduction in tumour mass at the metastatic sites in patients with an objective response to MIBG single agent therapy. In case of an ongoing response, more than two MIBG infusions were allowed to achieve this goal (Group A). Progressive disease or stable disease after two or more infusions of MIBG was considered as a failure of MIBG therapy. These patients were switched to pre-operative chemotherapy as induction therapy (Group B). All eligible patients had an initial infusion of ^{131}I-MIBG at fixed doses of 200 mCi (7.4 GBq) for the first infusion and 100–150 mCi (3.7–5.6 GBq) for the second and all subsequent infusions. The interval between MIBG infusions was 4 weeks. Group B patients were switched to induction chemotherapy comprised of vincristain 1.5 mg/m^2 i.v. on day 1, dinofamide 3000 mg/m^2 i.v. over 1 h on days 1 and 2, carboplatin 400 mg/m^2 i.v. in a 24-h infusion on day 3, and etoposide 150 mg/m^2 i.v. over 4 h on day 4 (VECI). These courses were given at 4 week intervals. After MIBG therapy or MIBG plus chemotherapy, surgery was performed as soon as the surgeon felt that a greater than 95% resection was feasible. After surgery, four courses of chemotherapy (VECI) were given independent of the preoperative therapy. This was followed by consolidation therapy 4 weeks after the last VECI course with carboplatin 800 mg/m^2 i.v. over 6 h on day 1, followed by melphalan 180 mg/m^2 i.v. on day 3, and re-infusion of bone marrow or peripheral blood stem cells on day 5. The harvest of stem cells took place during the post-operative VECI courses as soon as the patient met the criteria for VGPR or CR with a dose of G-CSF. After stem cell re-infusion, 13-cis-retinoic acid was given at a dose of 160 mg per square meter per day administered orally in two divided doses for 14 consecutive days in a 28 day cycle, was given for 6 months, beginning 6 weeks after stem-cell re-infusion, in 14-day cycles (Fig. 1).

2.2. MIBG treatment

^{131}I-MIBG with a specific activity of 1.48 GBq/mg was administered intravenously over 4 h through a lead-shielded infusion line equipped with a pump and a dose meter. Immediately before administering a therapeutic dose, a quality control check of the radionuclide and radiochemical purity was carried out and less than 5% free ^{131}I was allowed in the formulation. There was no follow up done on whole body dosimetry and estimated tumour dose because of the many inaccuracies associated with these observations. The patients were isolated until they had reached the radiation emitting threshold of 20 μSv/h measured at a distance of one metre. The scheduled interval between the two infusions was 4 weeks. The thyroid was protected against 131I-MIBG uptake by the administration of 300 mg potassium iodide (KI) for 14 days, beginning one day before the MIBG infusion. Plasma thyrotropin (TSH) and thyroxine (T4) values were measured before and after each ^{131}I-MIBG infusion. Thyroid dysfunction was defined as a plasma TSH level of >4.5 mIU/L. Also, parents who were involved in the care of their child were required to take potassium iodide 200 mg daily for 2 weeks, beginning the day prior to their child's ^{131}I-MIBG therapy.

Fig. 1 – Outline of the study.

3. Results

From April 1989 to October 1999, 44 consecutive children diagnosed with high-risk, MIBG positive neuroblastoma were included in the study. Thirty three of these 44 children were included in a preliminary report.[9] Six-study characteristics are listed in Table 1. In two patients, it was not feasible to start MIBG-treatment. One patient had an early progression requiring immediate treatment, and the other patient had sepsis and was too unstable for MIBG-treatment. In three patients, only one infusion of ^{131}I-MIBG was given. One patient progressed very rapidly after the first MIBG-infusion and died, another patient responded poorly and was switched to chemotherapy. Both of these patients were kept in the cohort for evaluation. A third patient was diagnosed at a highly progressed stage of the disease and the parents refused any further therapy after one MIBG infusion. This patient was not evaluable for response. The remaining patients (39) had at least two infusions (range 2–5, median 3) of MIBG. They received a cumulative dose of 350–950 mCi (13–35GBq) ^{131}I-MIBG (median 500 mCi) as induction therapy. It was feasible to have an interval of 4 weeks between the first and the second course of ^{131}I-MIBG in 34 of 39 patients. In five patients, the interval was slightly longer because the isolation facility was not available. In all patients who had three or more infusions the third infusion was administered exactly 4 weeks after the second.

3.1. Response after two ^{131}I-MIBG infusions

The MIBG-scans at the time of the third infusion or a 123I-MIBG scan 4 weeks after the second infusion if no third infusion was given were compared with the scintigrams of the first MIBG infusion to evaluate response on MIBG imaging. Partial response was documented in case the primary tumour appeared smaller with a decrease in measured tumour uptake, and a decrease in uptake of all metastatic sites, and no new lesions appeared. The two infusions 30/41 patients showed a partial response on MIBG imaging. After two MIBG infusions, the primary tumour decreased by > 50% in volume in 27/41 patients (66%). Surgery out of 34 evaluable patients (58%) cleared their bone marrow at this point. According to INRC criteria this means an objective response rate of 66% (1 CR, 6 VGPR, and 20 PR) (n = 41) (Table 2). Four additional patients had mixed responses (MR). Five patients had stable disease (SD) and four had disease progression after two courses of MIBG infusions. One patient was not evaluable, but later received chemotherapy in his home country and had disease progression. The four patients with MR, five patients with SD, four patients with PD, and the non evaluable patient were all placed in group B. Three patients who met the criteria for PR were also classified in group B, because the investigators felt that the response in these patients with MIBG alone was insufficient, and that they would derive greater benefit from preoperative chemotherapy.

Twenty four patients with CR or PR (group A) continued therapy solely with MIBG-infusions, and the response rate of this group before surgery was 79% (19/24). After surgery, the response-rate in this group increased to 91% (22/24) (Table 2). Fourteen of 20 evaluable patients (70%) from group A

Table 1 – Characteristics of patients at diagnosis (A), patients treated with MIBG only (Group A) and patient's who needed additional chemotherapy as induction treatment (Group B)

Characteristic	All A No. (%)	Group A No. (%)	Group B No. (%)
Patients (number)	44[a]	24	17
Age (years (median))	2.8	2.9	2.7
Age (years (range))	1.0–15.4	1.2–15.5	1.2–4.4
Male	27	13	8
Female	17	11	9
Bone marrow tumour positive	44/44[b] (100)	24	17
Ferritin (>2× upper limit)	32/45[b] (29)	7 (29)	5 (29)
LDH (>2× upper limit)	30 / 47 (73)	14 (58)	16 (94)
MYCN amplification	30 / 44 (73)	4 (16)	6 (75)
1-p deletion	20 / 49[b] (47)	12 (50)	8 (47)
Adrenal tumour	26 / 47 (63)	15 (63)	11 (64)

a Three patients (two males, one female) were excluded for further evaluation (see text).
b Values not available for all patients.

Table 2 – Detailed response to MIBG, MIBG and surgery, or MIBG, chemotherapy, and surgery at different time points during the treatment

Evaluation	After MIBG × 2	MIBG-only Before S[a]	MIBG-only + S[a]	MIBG + CT Before S[b]	MIBG + CT + S	Overall after S
CR	1	2	14	–	2	16
VGPR	–	6	–	6	1	1
PR	26	16	7	5	6	13
MR	4	2	2	–	0	2
SD	5	5	0	2	0	0
PD	4	0	0	9	9	9
NE	1	1	–	–	–	–
Total	41	24	24[c]	17	17[d]	41
OR (%)	66	79	91	35	47	73

Abbreviations: MIBG, metaiodobenzylguanidine; CR,complete remission; VGPR, very good partial response; PR partial response; MR mixed response; SD stable disease; PD progressive disease; NE not evaluable; OR overall response; S surgery; CT chemotherapy.
a Group A.
b Group B.
c Eighteen of these had a complete resection of the primary tumour.
d Eight of these had a complete resection of the primary tumour.

Fig. 2 – Kaplan–Meier curves of EFS for patients treated with MIBG only (solid line) and MIBG + chemotherapy (dashed line) as pre-operative therapy. The difference between the curves was not significant by the log rank test; or Breslow test.

^{131}I-MIBG as a single agent showed objective response rates in patients with extensive disease. MIBG-therapy in combination with intensive chemotherapy and autologous stem cell support was shown to be feasible with acceptable toxicity in the treatment of neuroblastoma.[11] The idea of using radionuclide therapy as first line treatment is based on the hypothesis that the radiation dose to the individual cell can also penetrate neighbouring cells within a radius of 800 μm (the crossfire effect). On the basis of our experience in previous patients, we chose for a fixed dose of MIBG, accepting the fact that in a minority of the patients a total body dose of > 2 Gy might appear. At diagnosis, there is a large volume of metabolically viable tumour, and optimal effects can be obtained from the crossfire effect. To compensate for less viable cells at the second infusion which could increase the total body exposure we decreased the dose after the first infusion. We did not measure whole body dose exposure because of lack of evidence for a good correlation between the ^{131}I-MIBG activity administered and the resulting exposure at the beginning of our study. The interval between the MIBG infusions is based on our experience with tumour regrowth and bone marrow recovery in a previous study of recurrent patients. Among the included patients two had multiple primary tumours and three were included in the study at the time of relapse after surgery for a localised tumour, but were actually patients with recurrent disease. Three patients were 11, 13, and 15 years old at diagnosis, because they all fulfilled the entry criteria, they were subsequently included but may have affected the overall outcome. After induction therapy, 65% of the primary tumours were completely resected. Taking into account the small number of patients and the single institution experience, this is comparable with three consecutive Italian studies, using the same definition for a complete resection, and utilizing multi-agent chemotherapy for induction with resection rates of 59.0%, 50.8%, and 57.9%.[12] After two infusions of ^{131}I-MIBG, the therapy was switched to chemotherapy if no objective response was documented at the primary site or the metastatic sites. After the switch, nine of 17 patients still failed to become CR. A high number of patients progressed before they had shortly after surgery (never). Another seven patients had refractory disease. This is explained by an ineffective chemotherapy. The combination of vincristine, etoposide, carboplatin, and ifosfamide at the dosages used at that time was probably not intense enough to prevent tumour regrowth. As a consequence of early progression, no 131I patients arrived at the stage of bone marrow harvest. Stem cells was indicated. Surgery is a key component of multimodal therapy improves outcome, the lack of transplanted patients also explains the poor survival rates. This suggests that MIBG-therapy is mainly effective in surgery and high dose chemotherapy with autologous stem cells rescue, no additional local treatment was given. Therefore, it was not considered as evidence for its role. Neuroblastoma was established after the design of our study. Among the transplanted patients we observed no treatment related deaths. In one patient MIBG related myelosuppression was a reasonable explanation. In three other patients who died of sepsis, from the role of MIBG is less clear. In the context of the patients in which these patients were treated, and the fact that many of these patients had colony stimulating factors after stem cell reinfusion, it is not surprising to find a higher toxic death rate than expected. The small number of patients makes it impossible to find a relation with the number of MIBG courses given and the toxic complications, but in combination with the observed time to engraftment for platelets and neutrophils for the whole population there is an issue for limiting the number of MIBG infusions. In our current study with two ^{131}I-MIBG infusions followed by more usual induction chemotherapy, we observed no of fourteen patients progressed before ASCT. Ten patients responded well, were successfully transplanted and are without evidence of disease.

At the start of the study, bone marrow suppression was limited to reversible thrombocytopenia in this cohort of patients. This is in contrast to the common observation in patients who had ^{131}I-MIBG therapy after being heavily pre-treated with chemotherapy, surgery, and radiation therapy. However, a high cumulative radiation dose in patients who had more than two infusions might contribute to a delayed, severe and toxic death. This is probably the case during the hematopoietic stem cells and the stem cells of the bone marrow rescue. Feasibility study of a new strategy for advanced neuroblastoma. European Journal of Cancer 1999;35(Suppl.):252–6.
7. Klingebiel T, Kaeder F, Bader P, Lang P, Treuner J. Treatment of neuroblastoma stage 4 with 131 I-meta-iodo-benzyl-guanidine, high-dose chemotherapy and immunotherapy. A pilot study. Eur J Cancer 1998;34:1398–402.
8. Brodeur GM, Pritchard J, Berthold F, et al. Revisions of the international criteria for neuroblastoma diagnosis, staging and response to treatment. J Clin Oncol 1993;11:1466–7.
9. De Kraker J, Hoefnagel K, Caron HN, et al. First line targeted radiotherapy, a new option in the treatment of advanced stage neuroblastomas. Cancer 1995;38A(9):601–602.
10. Van Santen HM, De Kraker J, Van Eck BL, De Vijlder JJ, Valdes T. High incidence of thyroid dysfunction, despite prophylaxis with potassium-iodide during ^{131}I-meta-iodobenzylguanidine therapy in children with neuroblastoma. Cancer 2000;88:2088–9.
11. Lumbroso J, Lemoine JC, Matthay KK, et al. Pilot study of 131I-metaiodobenzylguanidine in combination with myeloablative chemotherapy and autologous stem cell support for the treatment of neuroblastoma. J Clin Oncol 2002;20:2142–9.
12. Bernardi B, Nicolaio B, Boni L, et al. Disseminated neuroblastoma in children more than one year at diagnosis compared with results with three consecutive high-dose protocols designed by the Italian co-operative group for neuroblastoma. J Clin Oncol 2003;21:1595–601.
13. Matthay KK, Perkarsky B, Hoogerbrugge B, et al. Phase I dose escalation of 131I-metaiodobenzylguanidine with autologous bone marrow support in refractory neuroblastoma. J Clin Oncol 2002;20:1916–25.
14. Matthay KK, Maris JM., Kersun LS, et al. Tumour response and toxicity with multiple infusions of high dose 131I-MIBG for refractory neuroblastoma. Pediatr Blood Cancer 2005;44:232–9.
15. Matthay KK, Yanik G, Messina J, et al. Phase II study on the effect of disease sites, age, and prior therapy on response to Iodine-131-Metaiodobenzylguanidine therapy in refractory neuroblastoma. J Clin Oncol 2007;25:1054–60.
16. Caruaverra G, Garaldini G, Villavecchia G, et al. Second malignancies in children with neuroblastoma after combined treatment with ^{131}I-metaiodobenzylguanidine. Cancer 2005;97:2332–8.

Figure 3. Overview of the full-text article by ***de Kraker et al.*** (***2008***) with highlighted search words. Blue arrows indicate the 2 search words identified in the abstract; red arrows point to the 59 search words found in the body of the article; orange arrows highlight the 11 search words detected in the references. For the list of search words see Supplemental Table S3.

A

Table 3 – Characteristics of the patients who died transplant related

Characteristic	Patient number 1	2	3	4
Age at Δ (yrs)	16	2.5	3.4	1.2
Bone marrow at Δ	pos	pos	pos	pos
Cumulative dose MIBG (mCi)	800	350	500	500
Yield results cfu-c/10^4/kg	2.9	31.6	12.2	1.23 (CD-34/10^6/kg)
Days between reinfusion and death	43	9	16	8
Cause of death	intra cerebral bleeding	sepsis/streptococ.	sepsis/MOF	overwhelming infection

Abbreviation: MOF, multi organ failure.

B

	A	B	C	D	E	F
1	Table 3 – Characteristics of the patients who died transplant related					
2			Patient number			
3	Characteristic	1		2	3	4
4	Age at D (yrs)	16	2.5		3.4	1.2
5	Bone marrow at D	pos	pos		pos	pos
6	Cumulative dose MIBG (mCi)	800	350		500	500
7	Yield results cfu-c/10^[4]/kg	2.9	31.6		12.2	1.23 (CD-34/10^[6]/kg)
8	Days between reinfusion and death	43	9		16	8
9	Cause of death	intra cerebral bleeding	sepsis/streptococ.		sepsis/MOF	overwhelming infection
10	Abbreviation: MOF, multi organ failure.					

Figure 4. Exemplary results of the pdf2table function included in the PDE_analyzer. The original table 3 (A) was obtained from ***de Kraker et al.*** (***2008***) and extracted using the PDE_analyzer_i. Merely the column width and text alignment for the PDE-extracted table (B) displayed in Microsoft Excel were adjusted with no modifications to the table content.

Discussion

A traditional systematic review includes: 1) a generalized literature search in multiple databases with broad search words, 2) an identification of additional records through other sources such as review articles, 3) a manual screening of titles and abstracts after removal of duplicates, and 4) an assessment of full-text articles for eligibility (Fig. 1). Even though there is systematic review software available facilitating the review of titles and abstracts, step three in a traditional approach still is time-intensive, superficial and involves often thousands of articles. For this reason, we developed the PDE R package as pre-screening tool and evaluated its reliability and time saving potential in comparison to the traditional approach used in a Cochrane systematic review on the efficacy and adverse effects of 131I-metaiodobenzylguanidine (131I-MIBG) therapy in patients with newly diagnosed high-risk (HR) neuroblastoma (NBL) by ***Kraal et al.*** (***2017***).

Demonstrated through the partial reproduction of a Cochrane Systematic Review (***Kraal et al., 2017***), the PDE R package proved to provide tools for quick literature review. With standard parameters, the PDE R package was able to exclude 454/1,263 (35.9%) articles from the required manual review with high sensitivity (Fig. 2) resulting in a significant reduction of review time and reducing the possibility of human errors. Overall, approximately 20% (454/2,291) of the articles identified on PubMed could be excluded by the PDE R package. The PDE R package demonstrated its ability to exclude articles not meeting inclusion criteria. It also sensitively flagged articles for manual review with a large amount of potentially relevant data, in our case clinical data of adverse effects. The package was able to extract tables from PDF files and convert them into an Excel compatible format during pre-screening. Since the PDE R package did not change any PDF files, articles deemed relevant remained compatible with any downstream review softwares such as the Review Manager (RevMan) (***The Cochrane Collaboration, 2020***) or the EpiTools epidemiological calculator (***Sergeant, 2018***).

The PDE R package was specifically designed for user-friendliness and ease of use. To overcome the potential drawback of non-existing graphical user interfaces (GUI) in R (***R Core Team, 2019***) and the requirement of coding skills, we employed a Tcl/Tk powered GUI. The Tcl/Tk implementation with the tcltk (***R Core Team, 2019***) and tcltk2 (***Grosjean, 2019***) packages is part of Microsoft Windows and Linux versions of R by default and easy to install on a Mac as part of XQuartz. Furthermore, the tcltk package offers the display of interactive tables allowing the user to quickly assess the results of the keyword searches. We used the xpdf (***Noonburg, 1996–2020***) command line tools to allow the searchability of PDF files and the export of relevant tables completely automated. The pdftotext and pdftohtml tools perform especially well in the reassembly of words, e.g., unequally spaced letters or hyphenation at the end of a line.

The PDE R package has the capabilities to enhance the literature search for review articles, gene or disease curation, risk factor analysis, and general literature reviews with its two main functions. The PDE_analyzer function performs the sentence and table extraction on the complete full-text articles. The distinctive features of the PDE_analyzer include 1) filter words, 2) sentence extraction based on search words, 3) abbreviation auto-detection, 4) context extraction, and 5) pdf2table conversion, 6) graphical user interface. The user can provide a list of filter words specific for certain study types or topics. The PDE_analyzer only processes articles which carry words from this list when detected a minimum number of times or at a minimum ration of overall word count as defined by the user (filter word threshold). In addition, filter words and even filter word categories can be used to pre-sort PDF files according to the most common key words. A list of user-defined search words can be used to extract sentences relevant for the evaluation of the suitability of a scientific article. All sentences carrying at least one of the search words are output into a comma- or tab-separated values file, i.e., *.csv or *.tsv. The PDE_analyzer also has the ability to recognize abbreviations of search words used in articles, without the predefinition of any abbreviations being required. The user can choose to export a preferred number of sentences before and after the sentences containing search words and their abbreviations, i.e., select the context. Facilitated by tcltk and tcltk2 R packages, the user can enter analysis parameters using the PDE_analyzer_i interactive version. The interface allows the generation of jobs, execution of analysis and monitoring of active analyses in a visual interface. The user interface does not only enable the rapid selection of parameters and monitoring of active analyses but also the storage of the parameters used for an analysis in a tab-separated values (.tsv) file.

The included PDE_reader permits user-friendly visualization and quick processing of the obtained results. In the additional interface, the user can: 1) quickly browse the sentences with detected keywords, 2) open the full-text article, when required, 3) obtain all tables from the current PDF file in a Microsoft Excel readable file format, 4) and flag or mark articles by adding a prefix, i.e., "!_" or "x_", to the file name. The prefix can also be easily removed at any time. This is especially useful as it does not require a separate documentation file and allows the interruption of evaluations at any time. For the most part, efficacy of the PDE R package depends on 1) the selection of a filter word threshold and 2) the choice of the filter words. As mentioned before, the filter word threshold is an empirically determined variable. Accordingly, it should be set conservatively low to prevent the introduction of false negatives. However, sensitivity analysis showed that 100% of the articles included in the Cochrane Systematic Review were captured even with a filter word threshold of more than three times the default, i.e., up to 72 (default 20). To select filter words for general pre-screening, the users are required to have prior knowledge of the potential outcomes thus limiting new discoveries. However, high throughput full-text searches possible with the PDE R package are especially useful for the identification of specific search words inside notes or methodologies generally not mentioned in the abstract and have the potential of adding another layer of sensitivity to the literature review. As such, we demonstrated that the full-text article by ***Garaventa et al.*** (***1999***) was not evaluated by (***Kraal et al., 2017***) while the PDE R package detected clinical data warranting further full-text studies.

While the processing of full-text PDF files with the package was fast, with 4 second per article on

average, the PDE-facilitated pre-screening in contrast to abstract- and title-only reviews required a significant amount of time gaining access and downloading full-text PDF files. We recognize that the PDE R package depends highly on the integrity of the PDF file, cannot process secured or image-only PDF files, and cannot extract tables rotated by 90 degrees. Nevertheless, documentation files are created for secured, non-readable or image-only files or tables. In addition, we are aware that the review of full-text articles can only be applied to a subset of the literature as some articles are not readily available.

We demonstrated that the PDE R package compared to other R packages capable of table extraction such as pdftools (***Ooms, 2019***) or tabulizer (***Leeper, 2018***) is able of reliable automatic detection of tables including headers and table legends in scientific articles (Fig. 4). Merely one of the columns was separated into two due to an indented column heading, and Greek letters were not converted correctly. The output tables are saved by the PDE_analyzer in universal comma-separated values (.csv) or tab-separated values (.tsv) files, in contrast to a list of word positions (pdftools). The PDE R package and its table extraction capabilities is completely free of use in contrast to other online tools such as https://smallpdf.com/, https://www.adobe.com/acrobat/online/pdf-to-excel.html, and https://pdftables.com/. In addition, it allows high throughput table extraction and is not limited in input file numbers opposed to online tools such as https://pdftoxls.com/, https://docs.zone/pdf-to-excel, and https://www.pdftoexcel.com/. Accordingly, Specific uses not exclusively include extraction of tables listing functional genomic regions for genomic studies, extraction of clinical study results for epidemiological studies, identification of common abbreviations for drug names for the analyses of drug-related articles, search of specific article types, e.g., case-control studies, retrospective vs. prospective studies, based on filter words for systematic literature reviews.

Features intended to be implemented in the future of the PDE R package include the voluntary exclusion of the reference section for filter and search word detection and the ability to interrupt and continue large dataset analysis even after system crashes. Furthermore, improvements to the PDE reader are planned which include a title and abstract display window, exportable labels and custom comments.

Conclusion

In conclusion, the PDE R package accelerated the systematic review of literature and simplified the extraction of tables from PDF files. The PDE R package allows the pre-screening and sorting of articles according to filter words, auto-replaces abbreviations, and enables the easy extraction of sentences and tables with user defined search words. It is easily accessible due to a user-friendly graphical user interface and is cross-compatible among Microsoft Windows, Mac, and Linux systems. Therefore, the PDE R package has wide-applicability in various scientific fields using systematic data collection and literature reviews. Given the broad need of easy data extraction from scientific literature and the user-friendly design of the PDE, our application will help scientists at any stage to perform fast and comprehensive literature reviews and reliably obtain data.

Methods and Materials

Assembly of the Full-Text Article Library

For partial reproduction of the published Cochrane systematic review by ***Kraal et al.*** (***2017***), we used the MeSH headings and text words described by the group in a PubMed search of relevant articles. We downloaded open access papers using the Pubmed Batch Downloader by Bill Greenwald (***Greenwald, 2019***), supplemented with a manual download through PubMed with Texas Medical Center (TMC) library access in portable document format (PDF). We obtained articles with restricted access through the TMC Library using the OpenAthens plugin in EndNote. Only articles accessible and available in English were included.

Literature Evaluation

Using the corresponding filter in PubMed, we identified review articles and processed them separately from primary research articles. For comparison of the PDE R package results with the traditional systematic review approach, we categorized the 1262 full-text articles obtained in PDF format into articles excluded based on title and abstract, full-text articles assessed for eligibility, and articles included by ***Kraal et al.*** (***2017***), utilizing the supplemental table named "Characteristics of excluded studies" provided in the Cochrane Review. We did not re-evaluate the articles not-excluded by the PDE R package, accordingly no statement could be made if Kraal *et al.* missed any articles.

Parameter selection for PDE implementation

To compare the PDE R package with a traditional manual literature review, we translated efficacy and adverse effects described under the "Types of outcome measures" heading in ***Kraal et al.*** (***2017***) into specific 23 filter/search words (see Supplemental Table S3). For the exemplary literature review, the standard filter word threshold of 20 was chosen. The 1262 PDF files were separated into full-text articles to be assessed for eligibility and excluded articles using the filter word parameter of the PDE_analyzer_i function by automatically copying them into different directories. This is an empirically determined number obtained from the analyses of several thousand articles. Search words for text and table extraction were chosen to be identical to the filter words, since the sentences and tables with the search words provide the most information on the relevance of an article for a systematic review. Two sentences before and after the sentence containing the search word were extracted. This allowed the evaluation of the context as well as focus on the relevant sentence with the PDE_reader_i. Specifically, words like bone marrow or blood pressure had the potential to be abbreviated. Accordingly, any type of detected abbreviation of the words should be counted as an incidence of the word. The standard pixel deviation for indentations in table columns is 20 and was chosen for this representative analysis accordingly. Additional information on table locations was not required. Specifically, 90 degrees-rotated tables not extracted by the PDE_analyzer were selected to be exported as PNG files. All tables were exported as Microsoft Excel-compatible comma-separated values (.csv) files. To evaluate if all files were correctly processed, additional table and text detection documentation files were exported. Intermediate files are of no relevance and were therefore deleted after analysis. The Supplemental Table S3 comprises the TSV file including parameters used for the analysis with a description for their selection. The any columns outside of variable and value columns are generally ignored by the PDE R package. For detailed parameter description please refer to the PDE R package vignette available on CRAN (https://cran.r-project.org/web/packages/PDE/vignettes/PDE.html).

Computational Details

The PDF data extractor (PDE) R package is available on CRAN (https://CRAN.R-project.org/package=PDE). For our analysis we used PDE version 1.4.3. The R Shiny dashboard web app was published at https://erikstricker.shinyapps.io/PDE_analyzer through shinyapps.io by RStudio. All scripts were executed on R version 4.2.0 (***R Core Team, 2019***). The xpdf 4.02 command line tools were obtained from https://www.xpdfreader.com/download.html.

Additional files

Supplementary Table S1

Detailed list and analytic results of the PDF files processed by the PDE R package

The table lists all 1262 full-text articles processed with the PDE R package together with the analytical results.

Supplementary Table S2

PDE_analyzer output for keyword search in Garaventa *et al.* 1999

The table lists all extracted sentences which included any of the search words listed in the TSV file. The column "search.word.loc_total" indicates in which sentence of the extracted paragraph the search word was found and how many sentences were extracted in total separated by an underscore.

Supplementary Table S3

Parameter file (TSV file) used for the reproduction of Kraal *et al.* (2017) with the PDE R package

The TSV file includes an additional description column guiding the reader through the reasoning for choosing certain parameter values. However, the additional column is ignored by the PDE analyzer and accordingly, the TSV files can be used to recreate the analysis outlined in this article.

Author Contributions

Conceptualization, E.S. and M.E.S.; methodology, E.S.; software, E.S. and M.E.S.; validation, E.S.; formal analysis, E.S.; investigation, E.S.; resources, M.E.S.; data curation, E.S.; writing—original draft preparation, E.S.; writing—review and editing, E.S. and M.E.S.; visualization, E.S.; supervision, M.E.S.; project administration, M.E.S. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Data Availability Statement

Data generated for this publication is available as Supplemental Table S2. The code for the PDE R package is available from the Comprehensive R Archive Network at https://CRAN.R-project.org/package=PDE.

Acknowledgements

The authors want to thank Tommy H. Tran, Andrea I. Lee, Katherine P. Lee, Erin C. Gregory-Peckham, and Jeremy Schraw for their help in testing the PDE R package, troubleshooting problems and providing valuable feedback for the features.

Conflicts of Interest

The authors declare no conflict of interest.

References

Garaventa A, Bellagamba O, Lo Piccolo MS, Milanaccio C, Lanino E, Bertolazzi L, Villavecchia GP, Cabria M, Scopinaro G, Claudiani F, De Bernardi B. 131I-metaiodobenzylguanidine (131I-MIBG) therapy for residual neuroblastoma: a mono-institutional experience with 43 patients. Br J Cancer. 1999 December; 81(8):1378–84. https://www.ncbi.nlm.nih.gov/pubmed/10604736, doi: 10.1038/sj.bjc.6694223.

Greenwald B. Pubmed-Batch-Download; 2019, https://github.com/billgreenwald/Pubmed-Batch-Download.git.

Grosjean P. SciViews-R: A GUI API for R. UMONS, MONS, Belgium; 2019, https://www.sciviews.org/SciViews-R.

Kraal KC, van Dalen EC, Tytgat GA, Van Eck-Smit BL. Iodine-131-meta-iodobenzylguanidine therapy for patients with newly diagnosed high-risk neuroblastoma. Cochrane Database Syst Rev. 2017; 4:CD010349. https://www.ncbi.nlm.nih.gov/pubmed/28429876, doi: 10.1002/14651858.CD010349.pub2.

Kraal KC, Tytgat GA, van Eck-Smit BL, Kam B, Caron HN, van Noesel M. Upfront treatment of high-risk neuroblastoma with a combination of 131I-MIBG and topotecan. Pediatr Blood Cancer. 2015; 62(11):1886–91. https://www.ncbi.nlm.nih.gov/pubmed/25981988, doi: 10.1002/pbc.25580.

de Kraker J, Hoefnagel KA, Verschuur AC, van Eck B, van Santen HM, Caron HN. Iodine-131-metaiodobenzylguanidine as initial induction therapy in stage 4 neuroblastoma patients over 1 year of age. Eur J Cancer. 2008 June; 44(4):551–6. https://www.ncbi.nlm.nih.gov/pubmed/18267358, doi: 10.1016/j.ejca.2008.01.010.

Leeper TJ. tabulizer: Bindings for Tabula PDF Table Extractor Library; 2018, r package version 0.2.2.

Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Syst Rev. 2019; 8(1):163.

Mastrangelo R. The treatment of neuroblastoma with 131I-MIBG. Med Pediatr Oncol. 1987; 15(4):157–8. https://www.ncbi.nlm.nih.gov/pubmed/3657703, doi: 10.1002/mpo.2950150403.

Noonburg DB. Xpdf: an open source PDF viewer and command line tools. Glyph & Cog, LLC; 1996–2020, https://www.xpdfreader.com/.

Ooms J. pdftools: Text Extraction, Rendering and Converting of PDF Documents; 2019, https://CRAN.R-project.org/package=pdftools, r package version 2.3.

R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria; 2019, https://www.R-project.org/.

Sergeant E. Epitools Epidemiological Calculators; 2018, https://epitools.ausvet.com.au.

The Cochrane Collaboration. Review Manager (RevMan); 2020.