Precision Medicine Search for Paediatric Oncology

Bevan Koopman[1], Tracey Wright[1], Natacha Omer[2], Veronica McCabe[3], Guido Zuccon[4]
bevan.koopman@csiro.au,tracey.wright@csiro.au,natacha.omer@health.qld.gov.au
veronica.mccabe@childrens.org.au,g.zuccon@uq.edu.au
[1]CSIRO, [2]Queensland Health, [3]Children's Hospital Foundation, [4]University of Queensland
Brisbane, Australia

ABSTRACT

We present a search engine aimed to help clinicians find targeted treatments for children with cancer. Childhood cancer is a leading cause of death and clinicians increasingly seek treatments that are tailored to an individual patient, particularly their tumour genetics. Finding treatments that are specific to paediatrics and match individual genetics is a real challenge amongst the vast and growing body of medical literature and clinical trials. We aim to help clinicians through a search system tailored to this problem.

The system retrieves PubMed articles and clinical trials. Entity extraction is done to highlight genes, drugs and cancers — three key information types clinicians care about. Query suggestion helps clinicians formulate otherwise difficult queries and results are presented as a knowledge graph to help result interpretability. The proposed system aims to both significantly reduce the effort of searching for targeted treatments and potentially find life saving treatments that may have otherwise been missed. Demo details at http://health-search.csiro.au/oscar/.

CCS CONCEPTS

• **Information systems** → **Information retrieval**.

KEYWORDS

precision medicine, medical information retrieval

ACM Reference Format:
Bevan Koopman[1], Tracey Wright[1], Natacha Omer[2], Veronica McCabe[3], Guido Zuccon[4]. 2021. Precision Medicine Search for Paediatric Oncology. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3404835.3462792

1 PROBLEM AND TARGET USERS

Cancer is the leading cause of death for children and adolescents worldwide [9] with 400,000 cases a year [15]. A cure is found in 80% of cases in high-income countries and 15–45% in low-income countries [9]. Early diagnosis and targeted treatment are the main criteria for success. Cancer in children has a strong genetic component [12, 14]; 10% of all children with cancer have a genetic predisposition [18]. Treatments are particularly successful when tailored to the specific genetics of the child's tumour — this is the principle of precision medicine [1].

While treatments tailored to a child's tumour genetics can save their life, finding that treatment represents a significant challenge. Treatments can be hidden in two sources: clinical trials (ongoing, rigorous experiments to test new treatments) and medical literature. There is an ever growing collection of both.[1] Special search tools are needed to find the targeted treatment needle in this over growing treatment haystack. This paper describes such a system.

The user is a paediatric oncologist. They will know what cancer the child has been diagnosed with and the results of the child's genetic test. Given these two pieces of information they will formulate a query to search for treatments. Treatments can be organised in a rough hierarchy of preference:

- Paediatric treatments are preferred to adult only treatments. (Although most of the evidence out there will be in the adult space.) Paediatric patients are generally divided in two categories: < 12 years old and 12–16 years old (> 16 being considered an adult).
- Clinical trials that are in their latter phases (early phases are only concerned with safety rather than efficacy; final phases obtain approval for general use).
- Treatments specific to the cancer type of the patient are preferred to either the general cancers or to other cancer types.
- Treatments that act upon the specific genes are of particular interest and importance.
- Meta-analyses considering multiple studies are preferred over single randomised control trials, which are preferred over observational studies, which are preferred over single case studies.

According to the above, the paediatric oncologist will begin searching for the preferred type of treatment but will move 'down' the hierarchy if no suitable treatments are found. The key issue from this process is how laborious it is: for a single patient it may require 8 hours of searching and reading results to fully explore all the treatment options.[2] Finding a single relevant document describing the right treatment can be life saving.

2 SYSTEM OVERVIEW

Figure 1 provides an overview and the general workflow of the system. Each step in the process is described below:

❶ The information need comes from a specific patient: a child with a specific cancer type. The unique genetics of the patient's tumour are provided to the clinician in a report.

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3462792

[1]At the time of writing, ClinicalTrials.gov lists 367,512 trials and PubMed contains more than 27 million journal articles.
[2]Personal experience of paediatric oncologist author N. Omer.

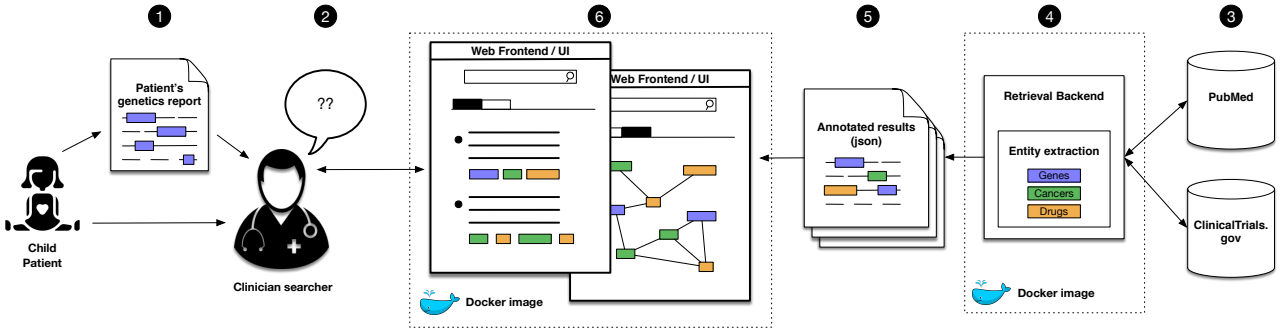

Figure 1: System overview and workflow.

❷ Given the cancer type and genetic information the clinician has to formulate a query to find targeted treatments.
❸ The system searches PubMed and ClinicalTrials.gov.
❹ All retrieved documents are annotated with the three entity types: genes, drugs and cancers.
❺ The annotated documents are returned to the frontend.
❻ The frontend renders the results, using the entity types to convey different facets to the clinician. Results can be viewed as a standard SERP and knowledge graph.

2.1 Entity Extraction

From a paediatric oncologists point of view, three types are information are important when searching for treatments: **genes**, which describe the unique characteristics of the patient; **drugs**, which in the cancer space are the main treatment type; and **cancers**, which is the specific type of cancer affecting the patient. We treat these as three entities types and develop our system around extracting mentions of these entities from documents (clinical trials and medical literature). Entity extraction is done using BERN: a neural, medical, named entity recognition tool [6]. BERN uses pre-trained BioBERT [10] to map free-text to biomedical entities. We employ and adapt BERN to output genes, drugs and cancers.

2.2 Retrieval Backend

Currently the retrieval backend acts as a meta search engine to two sources: PubMed and ClinicalTrials.gov. Both have API endpoints.[3]

The query is received by the retrieval backend and sent to the individual search components responsible for PubMed and ClinicalTrials. This is done in parallel to speed up response time.

On receipt of results from each service the result documents (trials or articles) are pushed onto a processing queue for the entity extraction pipeline. Multiple, parallel entity extraction 'workers' pop each document off the queue, annotate mentions of **genes**, **drugs** and **cancers**, then push the results onto a results queue. This allows for high throughout processing of entity extraction — a process that would otherwise be slow.

Once all results have been through entity extraction, they are passed to a reranker. Reranking preferences trials in latter phases and any documents relating to paediatrics. Finally, after reranking, the JSON results are passed back to the Web frontend.

[3]PubMed API: https://www.ncbi.nlm.nih.gov/home/develop/api/
ClinicalTrials.gov: https://clinicaltrials.gov/api/gui.

2.3 Web Frontend / UI

The clinician is presented with a free-text search box; results are shown in a SERP.[4] Figure 2 shows a screenshot of UI. The SERP is a mix of trials and PubMed articles; for each, a title and snippet is shown, plus associated entities displayed as tags or labels. Entities are colour-coded according to entity type (a legend for these is displayed below the search box). For clinical trials, a label shows the trial phase. Each document also has a widget indicating the age range to which this document refers (<12, 12–Adult, and Adult).

Results can be filtered using the sliders at the top of the results. Clinicians can narrow their results to: specific age ranges; certain clinical trail phases; and display only trials or PubMed articles.[5]

Finally, each document has a toggle button to save that document. There is a separate tab to view saved documents. This supports the clinician's workflow of spending time searching before then moving to the separate phase of reading documents in detail.

2.4 Entity-based Query Suggestion

Studies on how clinicians search have shown that query formulation is challenging [8]. There is high variance between clinicians, some being significantly better at formulating queries. These findings motivate the development of tools and techniques that aid the searcher with query formulation. One tried and tested method for this in information retrieval is query suggestion.

We use the entities contained in a set of search results as candidates for query suggestion. As the user enters a query, the system suggests matching entities from the retrieval results. (This methods obviously requires the user to have done an initial search to populate the result list.) Entities can be multi-word phrases and are colour coded according to their entity type. By clicking an entity, that entity is appended to the query. This allows the searcher to 1) reformulate their query according to the entities in their results; and 2) gain quick and easy insights into what entities are in their results. A sample query suggestion session is shown in Figure 3.

2.5 Knowledge Graph

While formulating effective queries in the medical domain is one issue, interpreting results is another. Even when presented with relevant results, clinicians may struggle to recognise them as such;

[4]The frontend is implemented in React and Bootstrap.
[5]The screenshot shows a check mark for 'Web'. This was to indicate that future work may incorporate results from the Web using, e.g., the Bing search API.

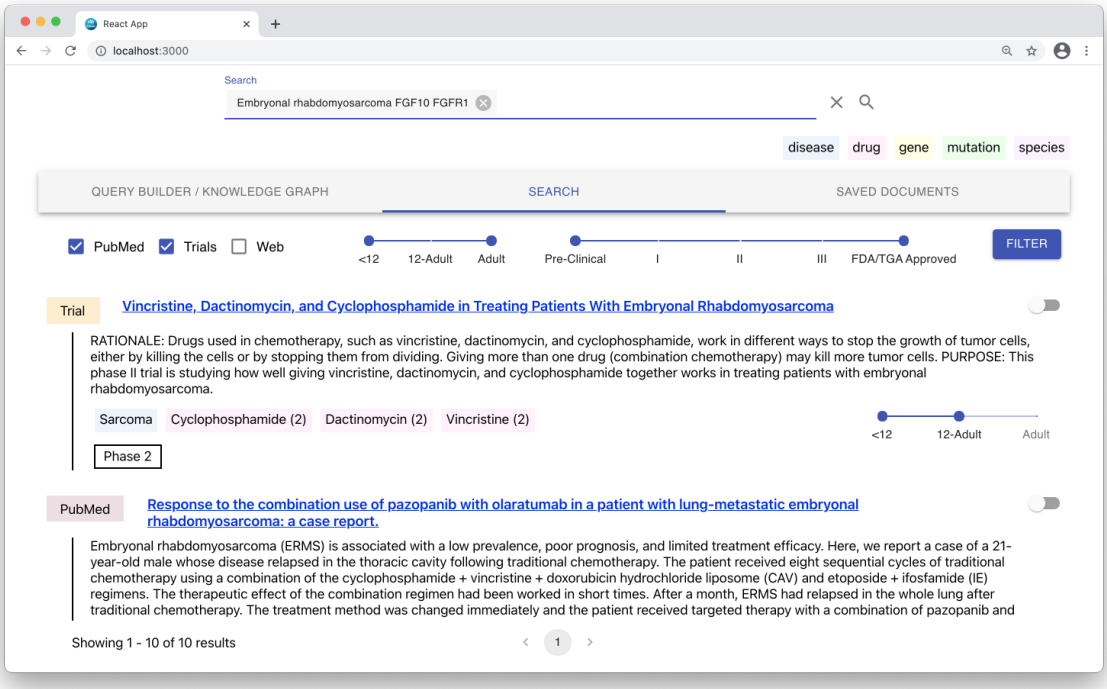

Figure 2: UI results. A sample SERP for the query '`Embryonal Rhabdomyosarcoma FGF10 FGFR1`'.

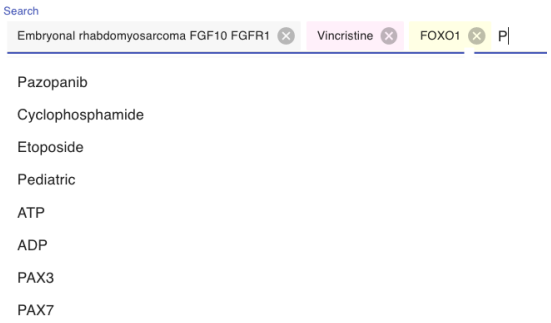

Figure 3: Sample query autocompletion session with three entities already added to the query. The clinician has typed 'P' and a list of suggested entities is provided.

this is because assessing relevance in the medical domain is non-trivial [7]. In fact, the ability for the searcher to interpret results can have a greater impact than the effectiveness of the IR system they use [17]. One issue being the difficulty to recognise relationships between disparate pieces of information [7]. One way to help make these relationships clearer is through a graph-based visualisation.

Toward this aim, we construct a knowledge graph of the search results. A sample knowledge graph for the query '`Embryonal Rhabdomyosarcoma FGF10 FGFR1`' is shown in Figure 4 (this is the knowledge graph for the results of Figure 2). Nodes in the graph represent the entities found in the set of search results. Edges currently represent co-occurrence of entities within some context window. Currently, the context window is co-occurrence within a medical article or clinical trial; however, this can be adapted to other context windows. In future the knowledge graph may integrate information from relevant external medical ontologies such as the Gene Ontology Resource [2]. The graph is interactive and can be manipulated by zooming or dragging nodes around to reorganise the graph. Right-clicking on a node/entity will display a pop-up that lists all the articles where that entity is mentioned. A small '+' button allows that entity to be appended to the query, thus allowing clinicians to refine their query while in the knowledge graph view.

3 COMPARISON WITH OTHER SYSTEMS

The system acts as a meta-search engine on top of ClinicalTrials.gov and PubMed. This was done for three reasons: enables us to reuse (and hence not reimplement) many of the base features that these systems support; it circumvents the needs to develop a separate component for updating a local index as more trials and literature gets added; it sits well with clinicians who are familiar with the use of PubMed and ClinicalTrials.gov. Using these two external services as the base search systems, we then focus on adding enhanced features that meet the paediatric oncology space: entity extraction based on genes, drugs and cancers; filtering based on age and trial phase; and a focus on query formulation and results visualisation.

Besides general search services such as PubMed and Clinicaltrials.gov, there are specialist services that help identify treatments. Apps such as UpToDate [5] provide evidence-based clinical information compiled and edited by medical professionals. While high quality, scope is very limited to common areas and well established treatments; thus it lacks both the breadth and recency requirements of paediatrics, rare cancers and access to the latest treatments.

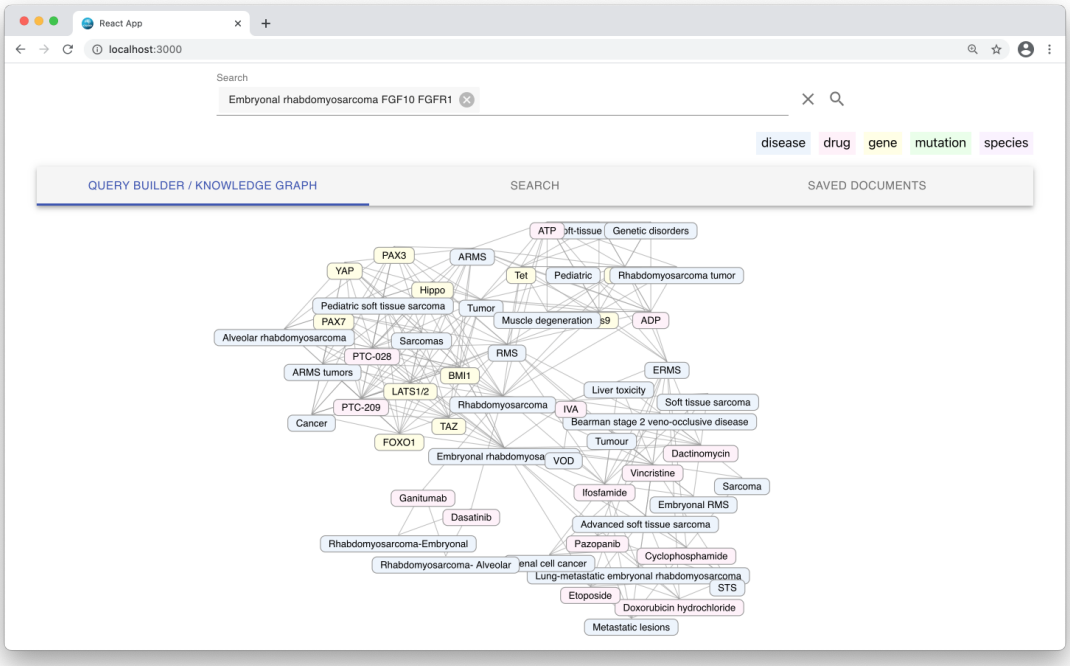

Figure 4: UI results. A sample knowledge graph for the query 'Embryonal Rhabdomyosarcoma FGF10 FGFR1'.

Specialist clinical trials recommender systems do cover the latest treatments and support clinician filters [3, 16]. However, such systems are nearly always specific to one medical area (e.g., haematology) and thus do not cover the breadth needed for paediatric oncology, nor the nuances of paediatrics. The systems also focus only on trials: they do not allow the clinician to search literature and have these results interleaved with trials within the same system.

4 IMPACT AND OUTLOOK

Currently our system does an initial retrieval (PubMed and ClinicalTrials.gov) and then reranks based on age and phase. Filters do provide some control for clinicians but the overall ranking can be improved. Specifically, ranking would be best achieved according the 'hierarchy of preference' presented in the introduction. This is a complex set of criteria so implementing a ranking methods that satisfies it will be equally complex. Learning-to-rank would be an attractive avenue to take. However, crafting the requisite training data is tricky — not only is compiling a sufficient volume laborious but the task itself makes training a single model difficult. This is because for some cases recall is paramount (e.g., rare cancers where it is essential to retrieve that one relevant treatment); while in other cases precision is the aim (e.g., for a common cancer that have a mix of relevant, partially relevant and non relevant results).

Good query formulation is hard in this domain where the clinician does not always know what treatments are out there. As such, efforts to aid query formulation should be pursued. We implemented an entity-based query suggestion method in this vein but more can be done. Query suggestions should account for the relationships between entities (e.g., via the knowledge graph). Previous work has developed visualisation tools for clinical queries [13]; this helps the clinician understand the impact of individual query terms on their results and we intent to incorporate this tool in future releases. Finally, query performance prediction methods could provide some quantitative feedback to the clinician about the efficacy of their query and could help guide effective query reformulation [4].

Knowledge graph construction is currently based on basic entity concurrence. There is a large body of research on knowledge graph construction that can be drawn upon to make this more effective [11]. In particular, the graph construction can be more contextualised to the current query and associated set of results. It is possible for the knowledge graph to become large and unwieldy so effective pruning or hiding techniques should be investigated.

The proof of the pudding is of course in the eating; thus all efforts should lead to an appropriately run user study with real patient cases and real paediatric oncologists. Evaluation will need to consider not only traditional IR effectiveness metrics but also effort or work saved because a paediatric oncologist's time is scarce and time spent searching is time not spent seeing patients.

The system described here aims to help paediatric oncologists find targeted treatments to childhood cancer. The system draws on two sources: medical literature from PubMed and trials from ClinicalTrials.gov. Entity extraction is done to capture three key types of information: drugs, cancers, genes. Result display is also based around these entity types, with an interactive, entity-based knowledge graph constructed, providing clinicians with a more interpretable alternative to a traditional SERP. Entity-based query suggestion is provided to help in query formulation.

A full fledged user study is planned to empirically evaluate the system with real patient cases and clinicians. If adopted in clinical practise, the system provides a necessary tool for precision medicine, tailoring the treatment to the individual. In the childhood cancer space, this is can be life-saving.

REFERENCES

[1] Euan A Ashley. 2016. Towards precision medicine. *Nature Reviews Genetics* 17, 9 (2016), 507.

[2] Seth Carbon, Eric Douglass, Benjamin M Good, Deepak R Unni, Nomi L Harris, Christopher J Mungall, Siddartha Basu, Rex L Chisholm, Robert J Dodson, Eric Hartline, et al. 2021. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research* 49, 1 (2021), 325–334.

[3] Maria Gonzalez, Matteo S. Carlino, Robert Richard Zielinski, Joel Smith, Robyn Saw, Angela Hong, Monika Keczkowska, Roslyn Ristuccia, Jim McBride, Alexander M. Menzies, and Georgina V. Long. 2016. An app to increase cross-referral and recruitment to melanoma clinical trials. *Journal of Clinical Oncology* 34 (2016), 9590–9590.

[4] Ben He and Iadh Ounis. 2006. Query performance prediction. *Information Systems* 31, 7 (2006), 585–594.

[5] Thomas Isaac, Jie Zheng, and Ashish Jha. 2012. Use of UpToDate and outcomes in US hospitals. *Journal of hospital medicine* 7, 2 (2012), 85–90.

[6] Donghyeon Kim, Jinhyuk Lee, Chan Ho So, Hwisang Jeon, Minbyul Jeong, Yonghwa Choi, Wonjin Yoon, Mujeen Sung, , and Jaewoo Kang. 2019. A Neural Named Entity Recognition and Multi-Type Normalization Tool for Biomedical Text Mining. *IEEE Access* 7 (2019), 73729–73740.

[7] Bevan Koopman and Guido Zuccon. 2014. Why Assessing Relevance in Medical IR is Demanding. In *MedIR at SIGIR*.

[8] Bevan Koopman, Guido Zuccon, and Peter Bruza. 2017. What Makes an Effective Clinical Query and Querier? *Journal of the Association for Information Science and Technology* 68, 11 (2017), 2557–2571.

[9] Catherine G Lam, Scott C Howard, Eric Bouffet, and Kathy Pritchard-Jones. 2019. Science and health for all children with cancer. *Science* 363, 6432 (2019), 1182–1186.

[10] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240.

[11] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. In *Conference on Empirical Methods in Natural Language Processing*. 3219–3232.

[12] Federica Saletta, Luciano Dalla Pozza, and Jennifer A Byrne. 2015. Genetic causes of cancer predisposition in children and adolescents. *Translational pediatrics* 4, 2 (2015), 67.

[13] Harrisen Scells and Guido Zuccon. 2018. Searchrefiner: A Query Visualisation and Understanding Tool for Systematic Reviews. In *Conference on Information and Knowledge Management (CIKM)*. 1939–1942.

[14] Logan G Spector, Nathan Pankratz, and Erin L Marcotte. 2015. Genetic and nongenetic risk factors for childhood cancer. *Pediatric Clinics* 62, 1 (2015), 11–25.

[15] Eva Steliarova-Foucher, Murielle Colombet, Lynn AG Ries, Florencia Moreno, Anastasia Dolya, Freddie Bray, Peter Hesseling, Hee Young Shin, Charles A Stiller, S Bouzbid, et al. 2017. International incidence of childhood cancer, 2001–10: a population-based registry study. *The Lancet Oncology* 18, 6 (2017), 719–731.

[16] Judith Trotman, Xavier Badoux, Admir Huseincehajic, Michele Gambrill, Anais LeGall, Michelle Daly, Mark Lacey, Sonia Byrne, Jennifer Aung, Shashi Nair, et al. 2013. Clintrial Refer-a Mobile App To Connect Patients With Local Clinical Trials. *Blood* 122, 21 (2013).

[17] Anton van der Vegt, Guido Zuccon, and Bevan Koopman. 2020. Do better search engines really equate to better clinical decisions? If not, why not? *Journal of the Association for Information Science and Technology* 72, 2 (2020), 141–155.

[18] Jinghui Zhang, Michael F Walsh, Gang Wu, Michael N Edmonson, Tanja A Gruber, John Easton, Dale Hedges, Xiaotu Ma, Xin Zhou, Donald A Yergeau, et al. 2015. Germline mutations in predisposition genes in pediatric cancer. *New England Journal of Medicine* 373, 24 (2015), 2336–2346.