Can Big Data Shed Light on the Origins of Pediatric Cancer?

A. Lindsay Frazier, MD, ScM,4 Mark Krailo, b Jen Poynter c

Twenty-five years ago, Sir Richard Doll and Sir Richard Peto estimated in their landmark monograph, *The Causes of Cancer*, that 75% to 80% of adult cancer could be attributed to lifestyle choices and environmental exposures.1 In stark contrast, elucidating the causes of childhood cancer remains one of the most vexing questions in medicine. Despite progress in identifying the molecular causes of many pediatric tumors, whether there are extrinsic environmental causes of pediatric cancer remains an elusive question. Without this evidence, the "primary" prevention of childhood cancer remains out of reach. The inherent epidemiologic challenge of studying the etiology of childhood cancer is simple math. Childhood cancer is a rare event, and the power to detect an association is limited by the small number of incident cases.

In this issue of *Pediatrics*, 2 companion articles by the same set of authors address this conundrum of rarity by using "Big Data" to address whether neonatal phototherapy can be implicated as a cause of childhood cancer. Wickremasinghe et al2 report the results of the California Late Impact of Phototherapy Study (CLIPS), a study of ~5 million infants who were identified by the California Office of Statewide Health Planning and Development by probabilistically linking vital statistics (birth and death certificates) with patient discharge data between 1998 and 2007. A total of 1100 infants (58 of whom had received phototherapy) developed cancer between 60 to 365 days of life. Because there are many factors previously identified that increase the risk of hyperbilirubinemia that might independently also affect cancer risk and thereby confound the association between phototherapy and infant cancer, the authors developed a “score” for each subject that captured all the reasons known to increase the need/propensity for phototherapy. This propensity score was necessary because it would not have been possible to adjust for each of these individual risk factors given the rarity of the outcome, infant cancer. The authors report, after propensity-adjustment, a significant association between receipt of phototherapy and increased risk of myeloid leukemia (adjusted odds ratio [aOR], 2.6; 95% confidence interval [CI]: 1.3–5.0) and kidney cancer (aOR, 2.5; 95% CI: 1.2–5.1) and “other” cancers (aOR, 1.6; 95% CI: 1.1–2.4) in infants exposed to phototherapy.

In the second paper, Newman et al3 report the results of the Late Impact of Getting Hyperbilirubinemia or Phototherapy (LIGHT) study, a retrospective cohort analysis that included 500 000 children born at Kaiser Permanente Northern California from 1995 to 2011. This study included all cases of childhood cancer and was not restricted to infants as in CLIPS. Overall, 711 children were diagnosed with cancer; 60 of these children had received phototherapy. Because this study was conducted in a single health care system, the actual individual covariates could be abstracted from the patient’s health record, including the bilirubin levels. Although the crude incidence rate ratios were elevated for myeloid leukemia and liver cancer ($P \leq .05$) and kidney cancer ($.05 < P \leq .10$),
after these rate ratios were adjusted with the propensity score, none of the associations remained significant. It should be noted that, as in CLIPS, the associations between phototherapy and cancer in the LIGHT study were stronger among children with Down syndrome.

But lest one be swayed by the allure of “Big Data,” even in these extremely large datasets, it is important to hone in on the actual number of cases observed. In the larger CLIPS study, the increased odds of myeloid leukemia were based on 10 cases in the exposed versus 103 in the unexposed. In the LIGHT study, the increased risk of liver cancer is based on 3 cases in the exposed versus 9 cases in the unexposed. Pediatric cancer remains a rare disease.

Although “Big Data” solves some problems, it creates others. The administrative datasets used in the Wickremasinghe paper were deidentified. The “probabilistic” linkage of these datasets means that the birth and death records were linked to the most likely admission and discharge codes, without certainty that it was truly the same patient. Because the actual individual is not identified, the report of cancer from the admission or discharge code cannot be cross-checked with another source, such as the California State Cancer Registry. This is unfortunate because the cancer diagnosis reported in the administrative datasets use International Classification of Diseases, Ninth Revision (ICD-9) codes rather than in the more precise childhood cancer codes of the International Childhood Cancer Classification. ICD-9 codes are usually abstracted by those responsible for billing, whereas International Childhood Cancer Classification codes used in cancer registries are assigned by specially trained cancer registrars. The potential for misclassification inherent in using the ICD-9 codes is apparent when one examines the cancer diagnoses contained in the “other cancer” category in CLIPS: bone cancer and skin cancer. Neither bone cancer nor skin cancer occurs in infants. These diagnoses must be misclassifications, likely of metastatic disease from other sites. An additional limitation in CLIPS is the lack of individual-level information, which makes it impossible to exclude the possibility of confounding by indication.

The design of the LIGHT study is an interesting counterbalance to the CLIPS study. Because this study took place within a single health care system, the diagnosis of cancer was derived from the actual hospital-based cancer registry, and the reports of the most prevalent cancers were further verified by medical record review. In addition, the authors could abstract the values of potential confounders, including total serum bilirubin, as well as construct a quasi-measure for the “dose” of phototherapy. The drawback to the LIGHT study is that the cohort, although still huge, is an order of magnitude smaller than the CLIPS study \( (5 \times 10^5 \text{ vs } 5 \times 10^6) \). Notably, after the more precise adjustment for the potential confounders is performed in the CLIPS study, the association of phototherapy with childhood cancer is no longer significant.

Study design aside, as practitioners, we are faced with deciding, based on these data, whether phototherapy causes infant and childhood cancers. Causal inference is, in the end, a judgment call. Famously, Bradford Hill suggested a set of criteria by which to judge causality that include: (1) temporality, (2) strength of the association, (3) dose-response, (4) reversibility, (5) consistency, (6) biologic plausibility, (7) specificity, and (8) analogy. If we apply these criteria to the data at hand, we see that several of these criteria have been met. The exposure is a plausible one; phototherapy is blue light, a type of ionizing radiation that has been shown to be associated with DNA damage. Most of the previous case-control studies have suggested an association between phototherapy and childhood cancer, principally acute myelogenous leukemia (AML) (consistency). Another form of ionizing radiation, UV light (analogy), is directly linked to the risk of skin cancer and is now classified as a type I carcinogen by the International Agency for Research on Cancer. In the LIGHT study, the association between AML and phototherapy did appear to be dose-related (biological gradient); however, this conclusion is based on 2 cases at the highest dose level. Although these data do support a potential role of phototherapy in childhood cancer to some degree, a number of questions remain. For example, why would phototherapy cause myeloid but not lymphocytic leukemia, and why kidney and liver cancers but not skin cancer (which is the cancer associated with UV radiation). These inconsistencies are difficult to fathom if phototherapy is indeed the cause of increased cancer incidence.

The authors further suggest that we should be concerned because of the prevalence of the exposure in the population and the potential attributable risk. In fact, based on data presented in the LIGHT study, the number of infants receiving phototherapy within the Kaiser Permanent Northern California health care system has increased from 2.7% of all children in 1995 to 15.9% of children in 2011, in part because of the introduction of home phototherapy. However, if the rates of phototherapy use are indeed increasing and causally related to the incidence of certain cancers (eg, AML, Wilms tumor, and hepatoblastoma), one might expect the incidence of these tumors to be increasing in the population in response to the increasing prevalence of the
exposure. However, several reports from population-based cancer registry data suggest that this is not the case.17–19

In the end, acknowledging that the information is imperfect, general pediatricians and neonatologists must make a choice. These data suggest that phototherapy may not be harmless, and that the risks as well as the benefits need to be weighed before flipping the switch.

**ABBREVIATIONS**

- AML: acute myelogenous leukemia
- aOR: adjusted odds ratio
- CI: confidence interval
- CLIPS: California Late Impact of Phototherapy Study
- ICD-9: International Classification of Diseases, Ninth Revision
- LIGHT: Late Impact of Getting Hyperbilirubinemia or Phototherapy

**REFERENCES**


Can Big Data Shed Light on the Origins of Pediatric Cancer?
A. Lindsay Frazier, Mark Krailo and Jen Poynter
_Pediatrics_; originally published online May 23, 2016;
DOI: 10.1542/peds.2016-0983

The online version of this article, along with updated information and services, is located on the World Wide Web at:
/content/early/2016/05/19/peds.2016-0983.full.html