Published in final edited form as:

Pharmacoepidemiol Drug Saf. 2012 May ; 21(Suppl 2): 37–43. doi:10.1002/pds.3241.

Merging of NCI-Funded Cooperative Oncology Group Data with an Administrative Data Source to Develop a More Effective Platform for Clinical Trial Analysis and Comparative Effectiveness Research: A Report from the Children's Oncology Group

R Aplenc^{1,2,*}, BT Fisher^{1,2,*}, YS Huang^{1,**}, Y Li^{1,**}, TA Alonzo³, RB Gerbing³, M Hall⁴, D Bertoach⁴, R Keren¹, AE Seif¹, L Sung⁵, PC Adamson¹, and A Gamis⁶

¹Children's Hospital of Philadelphia

²University of Pennsylvania DEClDE Center

³University of Southern California, Los Angeles

⁴Child Healthcare Corporation of America

⁵The Hospital for Sick Children, University of Toronto

⁶Children's Mercy Hospital and Clinics, Kansas City

Abstract

Purpose—National Cancer Institute (NCI)-funded cooperative oncology group trials have improved overall survival for children with cancer from 10% to 85% and have set standards of care for adults with malignancies. Despite these successes, cooperative oncology groups currently face substantial challenges. We are working to develop methods to improve the efficiency and effectiveness of these trials. Specifically, we merged data from the Children's Oncology Group (COG) and the Pediatric Health Information Systems (PHIS) to improve toxicity monitoring, estimate treatment-associated resource utilization and costs, and to address important clinical epidemiology questions.

Methods—COG and PHIS data on patients enrolled on a Phase III COG trial for *de novo* acute myeloid leukemia (AML) at 43 PHIS hospitals were merged using a probabilistic algorithm. Resource utilization summary statistics were then tabulated for the first chemotherapy course based on PHIS data.

Results—Of 416 patients enrolled on the Phase III COG trial at PHIS centers, 392 (94%) were successfully matched. Of these, 378 (96%) had inpatient PHIS data available beginning at the date of study enrollment. For these, daily blood product usage and anti-infective exposures were tabulated and standardized costs were described.

Conclusions—These data demonstrate that patients enrolled in a cooperative group oncology trial can be successfully identified in an administrative data set, and that supportive care resource utilization can be described. Further work is required to optimize the merging algorithm, map

Corresponding Author: Richard Aplenc, MD, PhD, Children's Hospital of Philadelphia, Colket Translational Research Building, 3501 Civic Center Boulevard, Room 4018, Philadelphia, PA 19104, 267-426-7252 (office), 267-725-0113 (fax), aplenc@email.chop.edu.

*Both authors contributed equally to this work

**Both authors contributed equally to this work

Conflict of Interest: The authors have no conflicts of interest to disclose.

resource utilization metrics to NCI Common Toxicity Criteria for monitoring toxicity, perform comparative effectiveness studies, and estimate the costs associated with protocol therapy.

Keywords

administrative data; acute myeloid leukemia; cooperative oncology group; comparative effectiveness; clinical trials

Introduction

National Cancer Institute (NCI)-funded cooperative group clinical trials have improved cure rates for children with cancer from less than 10% to approximately 85% and have set standards of care for the treatment of adult malignancies.[1] Despite these remarkable successes, NCI-funded cooperative group trials currently face substantial challenges and have important limitations. The importance of these challenges and limitations is clearly described in a recent Institute of Medicine Report (IOM): “The clinical trial system...is approaching a state of crisis. If [it] does not improve its efficiency and effectiveness, the introduction of new treatments for cancer will be delayed and patient lives will be lost unnecessarily.”[1]

While the list of challenges and limitations of the current clinical trial system is extensive, among the most pressing issues are the administrative burden of adverse event ascertainment, the absence of cost data, and the inability to perform comparative effectiveness research (CER) using available clinical trial data beyond analysis of the primary clinical trial question. Presently, adverse event reporting consumes more cooperative oncology group resources than any other activity.[2] Despite these expenditures, there are substantial concerns about both the accuracy and clinical relevance of the adverse event data. To our knowledge, no pediatric cooperative group trial has provided data on treatment costs or established patient cohorts that can serve as a platform for CER studies outside of specific, randomized interventions.

While administrative data sets can be used to perform CER studies and estimate treatment-related toxicities or treatment costs, such data sets have well-described limitations. These include uncertainty about actual patient diagnosis, difficulties in accurate risk stratification, an inability to determine specific treatment outcomes such as disease recurrence, and a paucity of malignancy phenotype data. Therefore, such data sets have had limited impact to date in pediatric cancer clinical epidemiology research.

Recognizing the importance of addressing the limitations in current cooperative group trials, we hypothesized that merging NCI-funded cooperative group data with hospital administrative data would create a dataset that would enable improved adverse event ascertainment, cost analysis, and the establishment of patient cohorts suitable for CER studies. We sought to test this hypothesis by merging data from the Children’s Oncology Group (COG) AAML0531 trial with administrative data from the Pediatric Health Information System (PHIS). While other investigators have merged data from adult NCI-funded cooperative oncology groups,[3–5] to our knowledge, administrative data have not previously been merged with pediatric cooperative group trial data.

Methods

Data Sources

Children’s Oncology Group (COG)—COG is the only NCI-funded pediatric cooperative oncology group and has approximately 240 participating centers in the United

States, Canada, Europe, and Australia (Figure 1). Annually, COG enrolls approximately 4,400 children on therapeutic oncology trials.[2] The AAML0531 trial enrolled 1,028 eligible patients on a randomized clinical trial of gemtuzumab with standard chemotherapy for the treatment of *de novo* acute myeloid leukemia (AML) from August 14, 2006 to June 15, 2010. Like all COG therapeutic trials, the AAML0531 trial collected extensive diagnostic, leukemia phenotype, and outcome data (Table 1). COG data are collected by Clinical Research Associates (CRAs) and submitted through an Internet-based remote data entry (RDE) system. A small number of data elements are submitted by central reviewers (leukemia morphology and cytogenetics) and laboratory personnel (leukemia molecular characteristics).

Pediatric Health Information System (PHIS)—PHIS currently includes data from 43 free-standing pediatric hospitals in the United States and captures approximately 85% of the free-standing pediatric hospitals registered with the Children's Hospital Association (CHA). As shown in Figure 1, PHIS sites represent the major metropolitan areas across the US. PHIS data have previously been used in over 100 peer-reviewed publications, including investigations of patients with AML.[6, 7]

PHIS data are drawn from inpatient units, emergency departments, observation units, and ambulatory surgery centers. The PHIS database includes the following: patient identification, demographics, dates of admission and discharge, International Classification of Diseases-9-Clinical Modification (ICD-9-CM) diagnosis (up to 41) and procedure (up to 41) codes, and specific billing/utilization data such as pharmaceuticals and blood products ordered, imaging requested, and clinical services utilized (Table 1). Associated with each medication and blood product order is the hospitalization day on which they were ordered and their route of administration. Member hospitals have access to the multi-center database via a secure web-based reporting system.

Oversight of PHIS data quality methods is a joint effort of the Child Health Corporation of America (CHCA; Shawnee Mission, KS; data management center), Thomson Reuters Healthcare (New York City, NY; data processing partner), and participating hospitals. After each hospital submits its quarterly data to Thomson Reuters, data are de-identified and data quality audits are performed. These audits primarily check for valid entries (e.g., valid ICD-9-CM diagnosis codes) and reasonable patient information (e.g., birthweight). Reports are generated that identify errors needing correction by the respective hospitals. Error rates above threshold values require hospitals to review their data and resubmit until error rates fall below the threshold values. Known data quality issues are transparently communicated to all PHIS data users. These data quality reports allow the data users to exclude data for quality reasons.

Data Merge and Statistical Analysis—COG statisticians created a data set of eligible patients enrolled on AAML0531 as of June 30, 2011. Patients in this data set were then matched in a probabilistic manner on treatment site, date of diagnosis, gender, and date of birth with a list of patients separately identified in PHIS by the presence of an ICD-9-CM code for AML (205.xx). Patients had to match on all criteria to be considered a valid match. The actual merging procedure was performed by CHCA, and the merged data set was then transferred by analysts at the Children's Hospital of Philadelphia (CHOP). Subsequent PHIS data-pulls for patients in the merged data set were performed at CHOP, as were subsequent data analyses. Standard summary statistics were used to summarize metrics of matching success and to tabulate resource utilization.

The standardized cost for each hospitalization was calculated using a cost master index that assigned a common cost across all hospitals to every item/service for which there is a CT

code in the PHIS database. The standardized cost for each CT code corresponds to the median of median hospital costs for that item. Costs for each item in each record in the PHIS databases were calculated by multiplying the charge for that item by the ratio of cost to charges (RCC) supplied by each hospital in each item category (pharmacy, laboratory, supplies, bed charges, other clinical services). The total standardized cost for each hospitalization was calculated by multiplying the number of units of each CT code by the standardized cost for that CT code, and summing up across all items/services in the billing record.

Protection of Human Subjects—All patients enrolled on AAML0531 gave informed consent for use of clinical trial data for research. All patient data remained de-identified throughout the merging and analytic process.

Results

Of 1,028 eligible patients enrolled on AAML0531, 416 (40%) were treated at institutions contributing to PHIS. Of those, 392 (94%) were successfully matched with a PHIS record. Of the 24 patients not identified, 10 were older than 18 and enrolled by pediatric oncology practices that typically admit adult patients to partnering adult institutions that do not submit data to PHIS. Eleven patients were enrolled at one of two centers, and the remaining three were enrolled at three different institutions.

Of the 392 successfully matched patients, 378 (96%) had PHIS data available for the first course of their AML chemotherapy. Six of the 14 patients without those data available had transferred from a non-PHIS institution after the first course of chemotherapy. Three patients were enrolled at one center, and the remaining five patients were missing without an identifiable cause. Match success rates for subsequent courses of chemotherapy ranged between 92% and 95%, as shown in Table 2.

Patients identified in PHIS did not differ from patients enrolled on AAML0531 in non-PHIS centers in regard to age, gender, or race distributions (Table 3). Although the percentage of Hispanic patients with PHIS data and those enrolled at non-PHIS sites was very similar, the percentage of Hispanic unmatched patients was strikingly low.

COG-PHIS merged data were used to determine resource utilization during the first course of chemotherapy. Among patients who survived induction therapy ($N = 369$), the average length of the induction course was 37.5 days (standard deviation/sd = 6.5, min = 3, max = 62). Patients spent an average of 29 days in the hospital during this first chemotherapy course (sd = 8.2, min = 1, max = 55). Blood product usage, specifically red cell, platelet, and clotting factor transfusions, is illustrated in Figure 2. Complete medication resource utilization data were available for 351 of the 378 patients. As an example of informative medication resource utilization data, Table 4 presents the number of anti-infective agents administered per 100 days of inpatient hospitalization. Of the anti-infective agents, antibiotics were most frequently administered at an average of 1.6 (95% confidence interval/CI 1.56 – 1.69) antibiotic exposures each hospital day.

The standardized cost master index was used to estimate standardized treatment costs for Induction 1 on AAML0531. The median standardized cost for Induction 1 was \$97,468 with no significant differences by gender, ethnicity, or treatment arm. However, treatment costs increased significantly by age categories of 0–1, 1–9, 10–19, and greater than 19 years: \$71,859, \$87,171, \$107,071, and \$197,614, respectively, $p = 0.0003$.

Conclusions

To our knowledge, this is the first report of merging pediatric cooperative oncology group and administrative data. Of the expected patients, 94% were appropriately person- and time-matched between the COG and PHIS data sources, and at least 92% of each patient's subsequent chemotherapy courses were accurately matched between the two data sources. Demographic characteristics were similar in patients at PHIS and non-PHIS sites and were similar to data reported for other COG trials in newly diagnosed AML patients[8] and in SEER-reported data.[9] These results demonstrate that PHIS and COG data can be successfully merged, and that the study population appears representative of pediatric AML patients who are not enrolled on COG trials or treated at non-PHIS sites.

The merged dataset provides a unique perspective on the supportive care required for successful AML therapy. The substantial blood product requirements are not unexpected, and the rapid decline in fresh frozen plasma and cryoprecipitate use and peak of platelet transfusion requirements at day 15 of treatment gives face validity to the blood product utilization data. The data on antimicrobial exposure also have face validity, since among pediatric AML patients, bacterial infections are more common than fungal infections, which in turn are more common than viral infections.[10] The standardized costs estimates are consistent with the only published study (single institution) of pediatric AML costs[11] and indicate that treatment costs associated with pediatric AML therapy are substantial.

Despite these strengths, the merged COG-PHIS data have important limitations. First, data are primarily limited to resources utilized in the inpatient setting. Thus, outpatient treatment complications will not be ascertained unless inpatient hospitalization is required for management. The impact of this will likely be modest for intensively treated pediatric malignancies, but may be more substantial for primarily outpatient-based treatments. Further work is required to define the extent of unascertainable events in a merged data set. In addition, data are not currently available on the generalizability of the COG-PHIS merged data set. Specifically, further work is required to determine whether patients in the COG-PHIS merged data set are truly representative of all pediatric patients with AML seen nationally in the US; whether merged patients are representative of all AML patients on COG trials; and whether merged patients treated on protocol therapy are similar to patients treated off-protocol. Finally, data are not yet available to determine whether patients in the merged data set receive supportive care similar to that given to patients outside the merged data set. We are actively seeking research funding to address all of these important questions.

Despite these limitations, a merged data set may be used to monitor toxicities on active clinical trials, perform CER studies, and execute cost analyses. Currently, toxicity reporting in cooperative group clinical trials depends on clinical research associates (CRAs) performing chart abstraction and discussing patients with clinical physicians.[12] However, this process has been shown to be variable in accuracy and completion[13] and consumes the greatest portion of COG resources.[2] Indeed, in order to reduce the burden of AE reporting, Kaiser et al. recently proposed reporting grade III/IV adverse events only in a subsample of patients on supplemental Phase III trials.[14] Although this option is attractive for reducing the burden on CRAs, it may result in missing data and will continue to require CRA resources.

COG-PHIS merged data may serve as an alternative resource for toxicity monitoring. While merged data will not completely replace CRA reporting of adverse events, some toxicities may be more efficiently and accurately monitored through the use of diagnosis codes, procedure codes, and pharmacy data contained in the merged data set. Work is ongoing to

identify which toxicities can be accurately monitored via the merged data to reduce the burden on CRAs. Ultimately, a combination of toxicity monitoring via merged data and CRA chart abstraction may result in both more accurate and more efficient toxicity reporting on NCI-funded cooperative group Phase III clinical trials.

The utility of merged data sets such as the COG-PHIS data is not limited to monitoring for adverse events. Creation of nationally representative homogeneous cohorts presents an opportunity to perform important CER studies particularly in supportive care therapies for chemotherapy complications. While multiple published studies have used hospital administrative data sets to complete CER studies, [6, 15, 16] critics of these publications appropriately cite established uncertainty of the cohort, both with regards to the disease and outcome of interest.[17] The COG-PHIS merged data sets address this concern directly, not only by creating a cohort of uniformly treated patients, but also by providing extensive data on patient and malignancy risk factors that can then be adjusted for in subsequent analyses.

These merged data provide benefits beyond improved confidence that the final cohort is truly representative of the disease of interest and utility in short-term CER studies. Specifically, these data may represent a novel opportunity for follow-up of patients after they have completed a chemotherapy trial. Additionally, these pediatric datasets could be linked to adult datasets so that children who have completed pediatric chemotherapy trials can be followed into adulthood for manifestations of late sequelae of treatment. The establishment of such large, prospective cohorts of long-term cancer survivors would likely be an invaluable resource for determining the longer-term burdens of cancer care.

Finally, the COG-PHIS merged data represent an important opportunity to perform cost analyses, to better understand resources utilized on a daily basis, and most importantly to evaluate the relationship between resource utilization and the quality and outcomes of care. To our knowledge, no pediatric Phase III oncology clinical trial has reported either the incremental differences in costs incurred from a randomized intervention or their relationship to quality and outcomes. This is a significant limitation of current randomized clinical trials, particularly given the increasing importance of demonstrating value for healthcare spending. In addition to comparing healthcare costs between the two study arms, analyzing data on resources utilized on a daily basis (such as blood product utilization, shown in Figure 2) may provide more specific information on the health care burden associated with a particular therapeutic intervention.

In summary, this study represents the first successful merger of data from a pediatric cooperative group oncology trial with data from a hospital administrative healthcare resource. The combined data draw on the strengths of the two individual datasets and afford the opportunity to perform toxicity monitoring, CER studies, and cost analysis research on a homogeneous patient cohort in an efficient manner. Furthermore, they establish a foundation for merging data from cooperative group trials of other pediatric malignancies with other available administrative databases.

Acknowledgments

The authors gratefully acknowledge Mr. Danniell Gaidula's graphic design contribution of Figure 1.

Merging of NCI-Funded Cooperative Oncology Group Data with an Administrative Data.

Research is supported by the Chair's Grant U10 CA98543-08 of the Children's Oncology Group from the National Cancer Institute, National Institutes of Health, Bethesda, MD, USA and NIH 1R01 CA133881 (Aplenc) The content is solely the responsibility of the authors and does not necessarily represent the official views of the NCI or the NIH.

A complete listing of grant support for research conducted by CCG and POG before initiation of the COG grant in 2003 is available online at: <http://www.childrensoncoogroup.org/admin/grantinfo.htm>

References

1. Nass, SJ.; Moses, HL.; Mendelsohn, J. A National Cancer Clinical Trials System for the 21st Century; Reinvigorating the NCI Cooperative Group Program. Washington, DC: Institute of Medicine; 2010. p. 317
2. Adamson P. 2011
3. Lamont EB, Herndon JE 2nd, Weeks JC, et al. Measuring clinically significant chemotherapy-related toxicities using Medicare claims from Cancer and Leukemia Group B (CALGB) trial participants. *Med Care*. 2008; 46(3):303–8. [PubMed: 18388845]
4. Lamont EB, Herndon JE 2nd, Weeks JC, et al. Measuring disease-free survival and cancer relapse using Medicare claims from CALGB breast cancer trial participants (companion to 9344). *J Natl Cancer Inst*. 2006; 98(18):1335–8. [PubMed: 16985253]
5. Lamont EB, Herndon JE 2nd, Weeks JC, et al. Criterion validity of Medicare chemotherapy claims in Cancer and Leukemia Group B breast and lung cancer trial participants. *J Natl Cancer Inst*. 2005; 97(14):1080–3. [PubMed: 16030306]
6. Fisher BT, Aplenc R, Localio R, et al. Cefepime and mortality in pediatric acute myelogenous leukemia: a retrospective cohort study. *Pediatr Infect Dis J*. 2009; 28(11):971–5. [PubMed: 19859014]
7. Fisher BT, Zaoutis TE, Leckerman KH, et al. Risk factors for renal failure in pediatric patients with acute myeloid leukemia: a retrospective cohort study. *Pediatr Blood Cancer*. 2010; 55(4):655–61. [PubMed: 20533519]
8. Cooper TM, Franklin J, Gerbing RB, et al. AAML03P1, a pilot study of the safety of gemtuzumab ozogamicin in combination with chemotherapy for newly diagnosed childhood acute myeloid leukemia: A report from the children's oncology group. *Cancer*. 2011
9. Ries, L.; Smith, M.; Gurney, J., et al., editors. *Cancer Incidence and Survival among Children and Adolescents: United States SEER Program 1975–1995*. National Cancer Institute; Bethesda, MD: 1999. p. 182
10. Sung L, Aplenc R, Zaoutis T, et al. Infections in pediatric acute myeloid leukemia: lessons learned and unresolved questions. *Pediatr Blood Cancer*. 2008; 51(4):458–60. [PubMed: 18561169]
11. Rosenman MB, Vik T, Hui SL, et al. Hospital resource utilization in childhood cancer. *J Pediatr Hematol Oncol*. 2005; 27(6):295–300. [PubMed: 15956880]
12. Mahoney MR, Sargent DJ, O'Connell MJ, et al. Dealing with a deluge of data: an assessment of adverse event data on North Central Cancer Treatment Group trials. *J Clin Oncol*. 2005; 23(36): 9275–81. [PubMed: 16361625]
13. Scharf O, Colevas AD. Adverse event reporting in publications compared with sponsor database for cancer clinical trials. *J Clin Oncol*. 2006; 24(24):3933–8. [PubMed: 16921045]
14. Kaiser LD, Melemed AS, Preston AJ, et al. Optimizing collection of adverse event data in cancer clinical trials supporting supplemental indications. *J Clin Oncol*. 28(34):5046–53. [PubMed: 20921453]
15. Zaoutis T, Localio AR, Leckerman K, et al. Prolonged intravenous therapy versus early transition to oral antimicrobial therapy for acute osteomyelitis in children. *Pediatrics*. 2009; 123(2):636–42. [PubMed: 19171632]
16. Newman K, Ponsky T, Kittle K, et al. Appendicitis 2000: variability in practice, outcomes, and resource utilization at thirty pediatric hospitals. *J Pediatr Surg*. 2003; 38(3):372–9. discussion 372–9. [PubMed: 12632352]
17. Benchimol EI, Manuel DG, To T, et al. Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. *J Clin Epidemiol*. 64(8):821–829. [PubMed: 21194889]

key points

The merging of NCI-funded cooperative oncology group clinical trial data with administrative data is possible. Such merged data sets facilitate clinical trial toxicity monitoring, comparative effectiveness studies, and analyses of protocol treatment costs.

Figure 1.
COG and PHIS Sites in the United States

Figure 2.
Daily Blood Product Administration during the First Chemotherapy Course

Table 1

Comparison of COG and PHIS Data

Clinical Covariates	COG	PHIS
Malignancy Diagnosis	Central Pathology Review	ICD9 Code
Data Sources	All Sites of Medical Care	Inpatient, Emergency Department, Observation Unit, Ambulatory Surgery
Risk Stratification Data	Extensive Clinical and Molecular Characterization	Procedure Billing Code without Laboratory Data
Treatment Response	Relapse, Death, Lost to Follow-up Explicitly Captured	Inpatient Vital Status (Death), ICD9 Code or Inferred from Procedure or Pharmacy Data(Relapse)
Treatment Toxicities	NCI Clinical Toxicity Criteria Reported by Data Managers	ICD9 Code or Inferred from Procedure or Pharmacy Data
Medications	No	Yes
Procedures	No	Yes
Blood Bank Resources	No	Yes
Radiology Services/Procedures	No	Yes (No Results)
Laboratory Services	No	Yes (No Results)
Cost Data	No	Yes

Table 2

Merge Success by Chemotherapy Course

	Total Patients	Identified
Course 1	416	378 (91%)
Course 2	359	341 (95%)
Course 3	321	304 (95%)
Course 4	226	210 (93%)
Course 5	190	175 (92%)

Table 3

Comparison of Age, Gender, Race, and Ethnicity for Patients Receiving Induction Therapy at PHIS and Non-PHIS Institutions

	PHIS (n=378)	Unmatched Patients* (n=38)	Non-PHIS (n=612)
Age			
<1 year	41 (11%)	5 (13%)	57 (9%)
1–9 years	163 (43%)	15 (39%)	241 (39%)
10–19 years	172 (46%)	14 (37%)	302 (49%)
> 19 years	2 (1%)	4 (11%)	12 (2%)
Gender			
Female	194 (51%)	17 (45%)	307 (50%)
Male	184 (49%)	21 (55%)	305 (50%)
Race			
Caucasian	270 (71%)	28 (74%)	454 (74%)
African-American	49 (13%)	5 (13%)	64 (11%)
Other/Unknown**	59 (16%)	5 (13%)	94 (15%)
Ethnicity			
Hispanic	80 (21%)	3 (8%)	106 (17%)
Non-Hispanic	288 (76%)	32 (84%)	480 (79%)
Unknown	10 (3%)	3(8%)	26 (4%)

* These patients were treated at PHIS contributing institutions but were not identified in the available PHIS data

** Demographic data based on COG data

Table 4

Mean number of antibiotic, antiviral, and antifungal exposure days per 100 hospital days during the first hospital admission for AML with 95% confidence intervals^{*}

Route of Administration	Antibiotics	Antiviral therapy	Antifungal therapy
Oral	28.8 (26.5–31.2)	7.3 (5.1–9.6)	41.0 (37.5–44.6)
Intravenous	134.1 (127.6–140.6)	5.9 (4.1–7.6)	29.4 (26.0–32.9)
Either oral or intravenous	162.9 (156.0–169.8)	13.2 (10.2–16.2)	70.4 (66.9–74.0)

^{*} Patients may have received more than one anti-infective on any given hospital day