Age-Specific Incidence of Acute Lymphoblastic Leukemia in U.S. Children: In Utero Initiation Model

Malcolm A. Smith, Timothy Chen, Richard Simon*

A large body of evidence supports the hypothesis that an infectious agent is involved in the etiology of acute lymphoblastic leukemia (ALL) in children, particularly those cases occurring in children between 2 and 5 years of age (1–4). Recently, a specific virus has been posited as a candidate etiologic agent for childhood ALL, with leukemia occurring as a consequence of primary infections of women during pregnancy that lead to in utero infection and to subsequent increased risk of developing ALL in early childhood (5). Since a critical component of this proposal is that in utero infection by a leukemia-inducing agent is the first step in development of many cases of childhood ALL, we were interested to see if a mathematical model with clinically reasonable parameters in which the critical leukemogenic event occurred in utero could be developed to explain the observed incidence of ALL among young children in the United States.

We examined models that represented the age at diagnosis of ALL as the sum of two time intervals. The first is the time from birth until the occurrence of an event that completes the neoplastic transformation in infants that was initiated in utero. The second time is the latent time from neoplastic transformation to clinical detection. Let the probability density function of the first time be denoted by \( f_1() \) and the probability density function of the latent time to clinical detection be denoted by \( f_2() \). We assume that these two time intervals are independent. Hence, the probability that leukemia is detected at time \( t \) in a child who experienced an initiating event prior to birth can be represented as

\[
\int_0^t f_1(z) f_2(t-z) dz,
\]

where the variable of integration \( z \) represents the random time of the transforming event.

For \( f_2 \), we used the normal distribution with mean \( \mu \) and standard deviation \( \sigma \). We examined two alternatives for \( f_1 \). Under the assumption that the risk of neoplastic transformation in an infected host is constant over time, \( f_1 \) should be an exponential distribution. Alternatively, it is possible that the number of infected cells that are susceptible targets for the second event is exponentially increasing over time from clonal expansion. Assuming that the clonal expansion is taking place at birth gives the limiting case and implies that \( f_1 \) should be the density of an extreme-value distribution. The hazard function \( h(t) \) for the extreme-value distribution is of the form \( h(t) = \rho_0 \rho_1 \exp(\rho_1 t) \), where \( \rho_0 \) and \( \rho_1 \) are parameters. The exponential distribution is a special case of the extreme-value distribution with \( \rho_1 = 0 \).

The model used contains three parameters in total when the exponential distribution is used for \( f_1 \) and four parameters in total when the extreme-value distribution is used. The parameters were estimated by fitting the models to the Surveillance, Epidemiology, and End Results (SEER) population-based age–incidence data (Fig. 1). In estimating the parameters, we made two modifications to the SEER data. We reduced the incidence rate for the first year of life from 28.8 to nine cases per 100,000, because approximately two thirds of these cases have genomic rearrangements involving the MLL gene at chromosome band 11q23 and are likely to have an etiology different from that of B-precursor ALL that occurs in children beyond infancy (6–8). We also corrected for T-cell ALL cases (see Fig. 1 legend), since although T-cell ALL represents only a small proportion of cases among children 2–3 years old, it represents 15%–20% of cases in children at the upper age range (9). The criterion used to fit the models to the data was minimal contingency chi-squared. Let \( O_i \) denote the observed proportion of cases that occurred in the \( i \)th year of life, for \( i = 1, \ldots, 10 \). Let \( E_i \) denote the expected value of this proportion predicted by the model. The model parameters were determined in order to minimize

\[
\sum_{i=1}^{10} \frac{(O_i - E_i)^2}{E_i}.
\]

Both models provided good fits to the data, as shown in Fig. 1. The three-parameter model gave an average absolute error of 10.2% and the four-parameter model gave an average absolute error of 10.4%. The estimated latency distributions were similar for both models (means of 1.6 and 1.7 years, respectively, and standard deviations of 0.7 year for both). For the exponential model, the exponential parameter is estimated as 0.30 (i.e., 30 events are expected for every 100 person-years of at-risk time). For the four-parameter model, the parameter values were \( \rho_0 = 0.30 \) (similar to the exponential parameter) and \( \rho_1 = 0.027 \). We also fit a model with a normally distributed latency distribution but with no second (i.e., postnatal) event required for leukemic transformation. We optimized the parameters of the latency distribution, but the model provided a very poor fit to the data with an average absolute error of 35%.

The fact that the four-parameter model did not fit the data better than the three-parameter model suggests that the target cell population for mutational events is not clonally expanding. In fact, the model can be modified to account for the possibility that the target population of cells is independently extinguished by a non-leukemogenic event whose time of occurrence is exponentially distributed with parameter \( \eta \). Under the assumption that the second event required for leukemia development oc-
the original model (with fied model is not distinguishable from initrating event is unobserved, the modi-
tional on the existence of the target
tion of the model is that children with an
leukemia, whereas the modified model
does not require that the in utero event
in a child who experiences an initiating event
 prior to birth is
\[ \int_0^\infty \delta e^{\delta z} e^{-\lambda z} f_2(t - z) dz, \]
which equates to
\[ \frac{\rho}{(\rho + \eta)} \int_0^\infty (\rho + \eta) e^{-\lambda (t + \beta)} f_2(t - z) dz. \]

The first factor of the integral is an exponential probability density. Since the number of children experiencing an initiating event is unobserved, the modi-
ified model is not distinguishable from the original model (with \( f_1 \) an exponential distribution) using age–incidence data. However, the biologic implications of the initial and modified models differ: The initial model predicts that all subjects experiencing the leukemia–initiating in utero event will develop leukemia, whereas the modified model does not require that the in utero event inevitably results in leukemia.

The models that we present are able to fit the observed age–incidence pattern for childhood ALL in the United States with parameters that are clinically rea-
sionable. Support for a latent period of approximately 1.5 years comes from the time course of relapse for children with B-precursor ALL, with recurrence of leukemia occurring not uncommonly at 3–5 years from diagnosis (i.e., 1–2 years from completion of therapy) (10,11).

A plausible physiologic interpreta-
tion of the model is that children with an in utero infection by a leukemia-
inducing agent may be placed in a pre-
leukemic state as a result of this infection. Those children in the preleukemic state at birth would then develop leukemia during childhood after a second event occurs. The second event might be mutation or loss of a critical tumor sup-
pressor gene or activation of an onco-
gene. It should be noted that the T anti-
gens of the polyomaviruses (e.g., simian virus 40 and JC virus) bind to p53 and induce genomic instability (12–17). Thus, the progression to the second event could be accelerated beyond nor-
mal rates by the biologic properties of the leukemia-inducing agent.

Little et al. (18) fit a variety of one-, two-, and three-mutation models to age–incidence distributions for ALL in En-
gland and Wales. They assumed that all events occurred after birth and did not account for a latent period between neo-
plastic transformation and clinical de-
tection. They found that one-event mod-
els did not provide adequate fits to the data. Two- and three-event models ade-
quately fit the data, but only if they assumed that the mutation rate re-
presenting the first event decreased abruptly after some age. Had they ac-
counted for a 1- to 2-year latent period in detection, the age at which the initial mutation rate decreases would likely have been very early, supporting our hy-
pothesis that the initial event occurs in utero.

In closing, the ability of the proposed model to explain the incidence curve of ALL in young children in the United States supports the hypothesis that a substantial number of cases of ALL among young children in developed countries such as the United States may be initiated by in utero events.

References

(4) Kinlen L. Evidence for an infectious cause of childhood leukaemia: comparison of a Scot-
(6) Ross JA, Potter JD, Robison LL. Infant leu-
ments in infant acute lymphoblastic leuke-


Notes

Editor’s note: SEER is a set of geographically defined, population-based central tumor registries in the United States, operated by local nonprofit organizations under contract to the National Cancer Institute (NCI). Each registry annually submits its cases to the NCI on a computer tape. These computer tapes are then edited by the NCI and made available for analysis.

We gratefully acknowledge the assistance of Lynn Ries in providing recent data for the incidence of childhood ALL from the SEER Program. More detailed incidence, mortality, and survival information will appear in a monograph dedicated to childhood cancers to be published by SEER.

Manuscript received April 10, 1997; revised July 31, 1997; accepted August 5, 1997.