International Journal of Computer Applications Technology and Research
Volume 10–Issue 12, 335 - 349, 2022, ISSN:-2319: 8656
DOI:10.7753/IJCATR1012.1008

Assessing Cancer Incidence Rates Across Age Groups Using Stratified Sampling and Survival Analysis Methods

Tahiru Mahama
Department of Statistics and
Actuarial Sciences,
Kwame Nkrumah University
of Science and Technology,
Kumasi, Ghana

Abstract: Cancer remains one of the leading causes of morbidity and mortality worldwide, with its incidence demonstrating marked variability across different age groups. Understanding these patterns is vital for effective prevention, diagnosis, and resource allocation. This study provides a comprehensive analysis of cancer incidence rates stratified by age, employing advanced statistical methodologies to uncover meaningful trends and survival outcomes. Beginning with a broad epidemiological overview, we analyze data from national cancer registries, establishing foundational insights into age-specific cancer prevalence and associated demographic patterns. Recognizing the heterogeneity in cancer distribution, the study applies stratified sampling techniques to ensure that each age group is proportionally represented, enhancing the reliability of comparative incidence analysis. Following stratified data extraction, we implement survival analysis methods, including Kaplan-Meier estimation and Cox proportional hazards modeling, to assess temporal patterns in patient survival across age strata. The analysis highlights age-related disparities in cancer prognosis, with younger cohorts generally exhibiting higher survival probabilities due to earlier diagnoses and more aggressive treatment regimens, while older cohorts face lower survival rates influenced by comorbidities and late-stage detection. Additionally, hazard ratios quantify the risk differential between age groups, offering granular insights into survival determinants. This integrated approach not only underscores the importance of age-specific analysis in oncological research but also provides actionable findings for public health policymakers, clinicians, and cancer epidemiologists. By combining stratified sampling with robust survival modeling, this study delivers a precise and nuanced understanding of how age influences cancer incidence and patient outcomes, contributing to more targeted screening strategies and personalized care protocols.

Keywords: Cancer incidence, Age stratification, Stratified sampling, Survival analysis, Kaplan-Meier estimation, Cox proportional hazards model

1. INTRODUCTION

1.1 Global Burden of Cancer and Demographic Transitions

Cancer remains one of the foremost public health challenges, contributing substantially to morbidity, mortality, and healthcare expenditure globally. As populations age and non-communicable diseases become increasingly prominent, cancer incidence and mortality rates have steadily risen across both high-income and low-to-middle-income countries [1]. This trend is driven by multifactorial elements including population aging, environmental exposures, changing lifestyle patterns, and improved case detection capabilities.

Demographic shifts—particularly increased life expectancy—have reshaped disease profiles, with more individuals surviving into age ranges where cancer risk is considerably elevated. Historically, infectious diseases dominated global health agendas; however, the growing dominance of chronic diseases such as cancer has led to a fundamental epidemiological transition [2]. Alongside urbanization and industrialization, this shift has increased the prevalence of risk factors such as tobacco use, sedentary behavior, and dietary changes, further compounding cancer risk [3].

In many settings, cancer now accounts for a larger proportion of premature deaths than communicable diseases. Health systems face escalating pressure not only to diagnose and treat cancers effectively but also to implement prevention strategies that target modifiable risk factors [4]. As survival improves due to early detection and therapeutic advances, population-level cancer control has become a cornerstone of public health planning.

The burden is not uniform. Different cancer types dominate in various regions due to genetic, behavioral, and environmental influences. For example, liver and stomach cancers remain more prevalent in East Asia, while breast and prostate cancers lead in Western countries [5]. Such variation underscores the need for region-specific interventions guided by robust surveillance systems and demographic-sensitive approaches.

To manage the growing cancer load, population-level data is indispensable for guiding resource allocation, screening strategies, and investment in infrastructure. Epidemiological modeling increasingly integrates these demographic variables to project future cancer trends with improved accuracy [6].

1.2 Importance of Age-Specific Cancer Incidence Assessment

Cancer incidence varies significantly by age, making age-specific analysis critical in understanding disease patterns and in crafting targeted prevention and control strategies. Most cancers demonstrate a steep rise in incidence after midlife, reflecting cumulative exposure to carcinogens, declining immune surveillance, and age-related genetic mutations [7]. Therefore, crude incidence rates may mask substantial heterogeneity within age groups, potentially leading to suboptimal policy decisions.

Age-specific incidence rates allow for nuanced interpretations of cancer trends by stratifying populations into relevant brackets—commonly five- or ten-year age bands. This granularity is essential for distinguishing between early-onset and late-onset cancers and for detecting anomalies that could signal underlying risk clusters or diagnostic shifts [8]. For instance, a rise in colorectal cancer incidence among adults under 50 may indicate emerging etiological concerns, whereas increases in those over 70 often reflect demographic expansion rather than epidemiologic transformation.

Adjusting for age is also vital when comparing cancer incidence between populations with different demographic structures. Standardized incidence ratios (SIRs) and age-standardized rates (ASRs) correct for these differences and enable valid cross-national or longitudinal comparisons [9]. Without such adjustment, younger populations might appear to have lower cancer burdens simply due to demographic makeup, not true epidemiologic variation.

Moreover, screening policies depend heavily on age-specific patterns. Mammography, colonoscopy, and PSA testing guidelines are informed by age-incidence curves that identify high-risk periods [10]. These age-related assessments enhance efficiency by ensuring that interventions target the segments most likely to benefit.

In sum, age-specific cancer assessment serves as both an analytical foundation and a practical guidepost for designing equitable, age-sensitive cancer control policies [11].

1.3 Role of Statistical Sampling and Survival Analysis in Modern Cancer Epidemiology

Robust cancer epidemiology relies on rigorous statistical sampling frameworks and survival analysis techniques to generate reliable, generalizable insights. Population-based cancer registries, health surveys, and hospital datasets form the backbone of epidemiologic inquiry, yet their utility hinges on representative sampling designs that minimize bias and allow accurate extrapolation [12].

Sampling methods such as stratified random sampling or multistage cluster sampling are commonly used in national cancer surveillance programs. These designs ensure inclusivity across age, sex, and geographic strata, enabling precise estimates of cancer incidence and prevalence at both national and subnational levels [13]. Moreover, they reduce standard errors and improve confidence interval precision, enhancing decision-making quality.

Survival analysis is a core component of cancer epidemiology. It evaluates time-to-event data, accounting for censored observations—cases where individuals have not yet experienced the event (e.g., death) at the end of follow-up. Techniques such as Kaplan-Meier estimation provide non-parametric survival functions, while Cox proportional hazards models allow for multivariable analysis and risk adjustment [14]. These methods have become indispensable for evaluating the impact of treatment, comorbidities, and socioeconomic status on cancer outcomes.

Importantly, survival data often complement incidence trends. A population may exhibit increasing cancer incidence but stable or improving survival due to earlier detection or better therapy. Thus, combining incidence and survival analysis yields a fuller picture of cancer control progress [15].

By leveraging representative sampling and advanced analytical tools, modern cancer epidemiology continues to evolve—supporting surveillance, policy design, and health equity monitoring at unprecedented levels of sophistication [16].

2. PRINCIPLES OF STRATIFIED SAMPLING IN EPIDEMIOLOGICAL RESEARCH

2.1 Sampling Theory and Advantages of Stratification

Sampling theory underpins much of population-based cancer epidemiology. Because it is often impractical to collect data from entire populations, researchers rely on samples that can yield valid, generalizable conclusions. In cancer surveillance, where timely and regionally relevant insights are required, probabilistic sampling remains the preferred method for reducing bias and enhancing representativeness [6].

A key advancement in this context is the use of **stratified sampling**, where the population is divided into non-overlapping subgroups, or strata, based on characteristics such as age, sex, or geography. Each stratum is then sampled independently, often with different sampling fractions. This approach ensures that smaller or more heterogeneous subgroups are adequately represented, increasing statistical power and enabling more granular analysis [7].

For cancer incidence and survival studies, stratification by **age group** is particularly useful. Cancer risk is not evenly distributed across age; older populations have substantially higher baseline risk. Without stratification, age-related variation can distort both point estimates and temporal trends. Stratified designs mitigate this by ensuring each age group is proportionally included in the sampling frame [8].

Moreover, stratification enables **post-stratification weighting**, improving the precision of estimates for national or regional cancer registries. This is especially valuable when conducting subsample analyses or investigating rare cancers, where unstratified designs may yield insufficient cases [9].

Stratified sampling also enhances **comparability across datasets** by facilitating standardization. Datasets from different regions or time periods can be more effectively harmonized if sampling procedures and age stratification criteria are aligned. This improves the utility of pooled analyses and meta-analyses in global cancer epidemiology [10].

In summary, stratified sampling is a foundational principle that strengthens the reliability, validity, and efficiency of population-based cancer studies by ensuring structured inclusion of diverse demographic groups.

2.2 Defining Age Strata in Cancer Studies: WHO and SEER Standards

The definition of age strata is a critical consideration in designing cancer epidemiologic studies. Age not only influences disease onset but also affects clinical presentation, treatment response, and survival outcomes. Accordingly, standardization of age categories enhances comparability across surveillance systems and analytical studies [11].

The World Health Organization (WHO) and the Surveillance, Epidemiology, and End Results (SEER) program offer widely accepted frameworks for age stratification. The WHO often recommends using five-year age bands beginning at birth (e.g., 0–4, 5–9, ..., 80–84, 85+), a structure that aligns with most global mortality and morbidity datasets. This system allows for high-resolution age-specific rate calculations and is especially effective for visualizing trends over time [12].

SEER, a major U.S.-based cancer registry system, typically adopts similar stratifications but sometimes aggregates age bands for specific analyses or rare cancers. For example, broader intervals such as 0–19, 20–44, 45–64, and 65+ may be employed to simplify presentation or enhance statistical stability in low-frequency strata [13]. Such flexibility ensures analytical efficiency while preserving epidemiologic relevance.

Selection of strata length can be influenced by **expected incidence distribution**, population size within each age group, and the research question at hand. Shorter intervals offer greater resolution but may introduce instability if sample sizes are small. Conversely, longer intervals may mask key patterns, such as early-onset cancers or age-specific screening effects [14].

Accurate age categorization is also essential for **age-standardized incidence rate (ASIR)** calculations, which adjust for population age differences. These rates are critical for cross-national and intertemporal comparisons. Misaligned age strata can compromise ASIR validity, reducing interpretability and misleading policymakers [15].

Thus, adherence to globally recognized age grouping standards facilitates methodological rigor and enhances data interoperability in cancer research settings.

2.3 Sample Allocation Techniques: Proportional vs Optimal Stratified Sampling

Once strata are defined, the next methodological decision involves determining how samples should be allocated across them. Two common strategies in stratified sampling for cancer epidemiology are proportional allocation and optimal (or Neyman) allocation, each with specific strengths and limitations [16].

Proportional allocation assigns sample sizes in each stratum according to the proportion of that stratum in the overall population. For example, if individuals aged 60–69 constitute 20% of the national population, then 20% of the survey sample would be drawn from this age group. This method is straightforward and aligns the sample structure closely with the source population, facilitating generalization of prevalence estimates [17].

While intuitive and widely used in large-scale cancer registries, proportional allocation can underrepresent important subpopulations—particularly younger or older age groups with relatively low population shares but high epidemiological interest. For example, early-onset cancers in younger adults may require deliberate oversampling to yield stable estimates [18].

In contrast, optimal (Neyman) allocation assigns sample sizes based on the standard deviation within each stratum and their relative sizes. Strata with greater internal variability are sampled more intensively to minimize overall variance of estimates. This is especially useful when the objective is to obtain precise estimates of incidence or survival with limited resources [19].

Optimal allocation, however, may distort the representativeness of the sample unless weights are applied during analysis. Post-stratification weighting can adjust for such disproportionality but requires careful implementation to avoid inflating standard errors or introducing bias [20].

In practice, a hybrid approach is often used. Health systems may begin with proportional allocation to mirror the general population but apply optimal allocation principles to oversample strata known to exhibit high variance or clinical relevance. This strategy ensures both generalizability and analytic power, particularly in multicenter or longitudinal cancer studies [21].

The choice of allocation method must also consider logistical constraints. Optimal designs often demand detailed prior knowledge of variance components, which may not be available for newly introduced cancers or underrepresented populations. In such cases, proportional allocation remains a robust default [22].

Ultimately, sample allocation strategies must align with the study’s epidemiologic objectives—whether prevalence estimation, risk factor identification, or survival modeling. Thoughtful application of allocation techniques strengthens statistical precision and enhances the policy relevance of cancer surveillance programs.

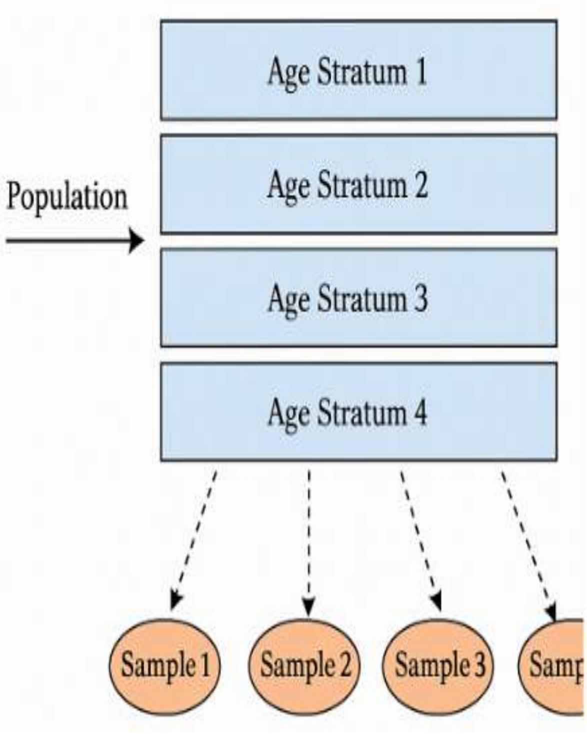

Figure 1: Diagram illustrating stratified sampling by age groups

Table 1: Stratification Schema Showing Population Weights and Sample Sizes by Age Group

Age Group (Years)	**Population Proportion (%)**	**Sample Weight**	**Sample Size (n = 10,000)**
0–14	18.5	0.185	1,850
15–24	14.0	0.140	1,400
25–44	26.5	0.265	2,650
45–64	24.0	0.240	2,400
65–74	9.0	0.090	900
75+	8.0	0.080	800

3. DATA SOURCES AND COHORT CHARACTERISTICS

3.1 Overview of Cancer Registry and Health Survey Data Used

Population-based cancer registries and health survey programs form the backbone of modern cancer epidemiology. Among the most extensively utilized sources are the **Surveillance,** Epidemiology, and End Results (SEER) program and the GLOBOCAN project. These platforms collect standardized data on cancer incidence, mortality, and survival across diverse geographic and demographic contexts [11].

SEER is notable for its granular, individual-level data from population-based registries, encompassing detailed information on tumor characteristics, treatment modalities, and patient demographics. The registry includes multiple geographic regions with demographic diversity, thereby supporting national-level estimates and trend analyses. SEER's methodological rigor and data quality controls have made it a global model for cancer registration [12].

GLOBOCAN, administered by the International Agency for Research on Cancer (IARC), provides aggregate-level cancer statistics for countries worldwide. It integrates information from national registries, vital statistics, and regional estimates to offer a comprehensive overview of global cancer burden. GLOBOCAN's focus on low- and middle-income countries helps bridge critical knowledge gaps in regions with limited surveillance infrastructure [13].

Other valuable datasets include national health interview surveys and health examination surveys, which provide self-reported data on cancer screening, risk factors, and access to care. While these lack the diagnostic confirmation of registry data, they offer rich contextual information for behavioral and policy research [14].

Integration of these data sources enhances analytical power and external validity. For example, linking SEER data with Medicare claims or survey responses allows researchers to explore treatment patterns, comorbidities, and quality-of-life outcomes in cancer populations [15]. Each dataset contributes distinct strengths, and their combined use supports a multifaceted understanding of cancer epidemiology across demographic and clinical dimensions.

3.2 Inclusion/Exclusion Criteria and Event Definitions

In cancer epidemiology, clear inclusion and exclusion criteria are essential for defining the study population and ensuring analytical consistency. Most registry-based studies limit inclusion to individuals with a confirmed primary cancer diagnosis, recorded using International Classification of Diseases for Oncology (ICD-O) codes. Typically, only first primary malignancies are included to avoid confounding from prior cancer history [16].

Age restrictions are commonly applied to align with population structure or study goals. For instance, individuals under age 20 may be excluded when the focus is on adult-onset cancers. Conversely, studies of pediatric malignancies apply upper age limits, often excluding patients over 19 years. Defining appropriate age thresholds allows for homogeneity within strata and ensures comparability with external datasets [17].

Event definitions depend on the analytic focus. In incidence studies, the event of interest is usually the date of cancer diagnosis, validated through histopathology, imaging, or clinical confirmation. In survival analyses, the event may be death from any cause or cancer-specific death, depending on the research question. Accurate event coding is vital for estimating time-to-event outcomes such as median survival or hazard ratios [18].

Exclusion criteria often address data completeness and quality. Records lacking critical variables (e.g., diagnosis date, tumor stage, or survival time) are typically omitted. Cases with missing sex or inconsistent age information may also be excluded to prevent misclassification. Additionally, data from sources with known underreporting or non-standard collection methods may be removed from pooled analyses [19].

To enhance reproducibility, eligibility criteria are pre-specified and documented alongside flow diagrams detailing the number of included and excluded cases. This transparency facilitates cross-study comparisons and ensures methodological rigor when using cancer registry or survey datasets [20].

3.3 Distribution of Demographic and Clinical Variables Across Age Strata

Understanding the distribution of demographic and clinical variables across age groups is crucial in interpreting cancer burden and identifying disparities. Age-stratified analyses enable researchers to uncover differences in incidence patterns, disease progression, and treatment access that would be obscured in aggregate-level summaries [21].

Demographic variables such as sex, race/ethnicity, and geographic location exhibit clear variations across age cohorts. For instance, younger age groups may show higher proportions of racial or ethnic minorities in urban areas, while older groups may be predominantly non-Hispanic White and more rural. These distributions are influenced by historical fertility trends, migration patterns, and regional population dynamics [22].

Educational attainment and socioeconomic status (SES) also vary with age. Older individuals may have lower average educational levels, especially in rural settings, which can influence access to screening and treatment. Younger patients, although better educated on average, may lack health insurance coverage, affecting early diagnosis and continuity of care [23].

Clinical characteristics such as tumor stage at diagnosis, cancer type, and treatment modality are strongly age-dependent. Pediatric and adolescent populations typically present with hematologic malignancies or bone tumors, while epithelial cancers dominate in older adults. Among those aged 65 and above, delayed diagnosis is more common, often resulting in later-stage presentation [24].

Treatment modality choices—surgery, radiation, chemotherapy—also differ. Younger patients may receive aggressive multimodal treatments due to better baseline health and treatment tolerance. In contrast, comorbidities in older adults may limit therapeutic options, influencing both prognosis and quality of life [25].

Analyzing these patterns requires thoughtful stratification and visualization techniques. Cross-tabulations, stacked bar plots, and age-specific prevalence curves help illustrate relationships between demographic factors and clinical indicators. Stratified regression models may also be employed to examine interaction effects, such as how socioeconomic status modifies the association between age and treatment receipt [26].

Such disaggregated analyses guide the development of targeted interventions. For example, identifying higher late-stage diagnosis rates in a specific age-ethnicity group can inform culturally tailored outreach and early detection programs. The age-stratified view not only refines epidemiologic understanding but also supports more equitable health system responses to cancer.

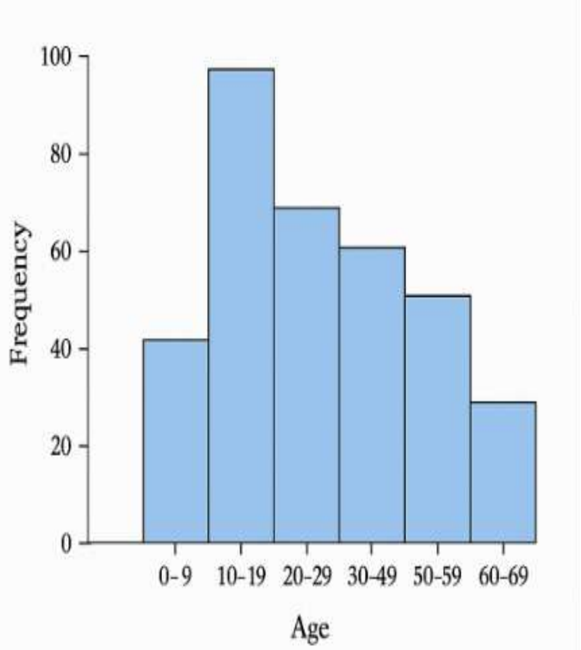

Figure 2: Histogram of age distribution in the study population

Table 2: Descriptive Statistics by Age Group

Age Group (Years)	Mean Age (Years)	Male (%)	Female (%)	Most Common Cancer Type	Prevalence (%)

Age Group (Years)	**Mean Age (Years)**	**Male (%)**	**Female (%)**	**Most Common Cancer Type**	**Prevalence (%)**
0–14	9.1	51.3	48.7	Leukemia	34.5
15–24	19.6	50.8	49.2	Hodgkin Lymphoma	27.1
25–44	35.2	48.9	51.1	Breast Cancer	29.8
45–64	54.7	49.6	50.4	Colorectal Cancer	22.4
65–74	69.1	52.0	48.0	Prostate Cancer (male)	35.7
75+	81.3	47.8	52.2	Lung Cancer	31.5

4. SURVIVAL ANALYSIS FRAMEWORK

4.1 Time-to-Event Data: Definitions of Survival Time, Censoring, and Event

Time-to-event data—commonly referred to as survival data—are central to cancer epidemiology. These data types capture the duration from a defined starting point to the occurrence of a specific event, typically death, recurrence, or disease progression. Unlike ordinary outcomes, survival data include a temporal component, making their analysis distinct from cross-sectional or point-prevalence measures [15].

The definition of survival time varies depending on study design. For most cancer studies, it is measured from the date of diagnosis to the occurrence of the event of interest. This metric may also begin from treatment initiation or randomization in clinical trials. Accurate recording of this time is crucial, as even small discrepancies can significantly impact statistical conclusions, particularly in short follow-up studies [16].

A critical feature of survival data is **censoring**. This occurs when the event of interest has not happened for a subject during the observation period. Right-censoring is the most common type, typically occurring when a patient is lost to follow-up or remains alive at the study's end. Proper handling of censored cases preserves information and prevents bias in survival estimation [17].

The **event** must be clearly defined, whether it is all-cause mortality, cancer-specific death, recurrence, or metastasis. These distinctions matter, especially in populations with high comorbidity burdens, where cause-specific survival may differ substantially from overall survival. Misclassification of events can distort survival curves and mislead interpretation [18].

Time-to-event data require specialized analytical techniques that respect the temporal structure and censoring mechanisms. Standard linear models are inappropriate, as they assume continuous or normally distributed outcomes and do not account for varying follow-up times. Thus, survival analysis offers tailored methodologies that accommodate these complexities while extracting meaningful clinical insights [19].

4.2 Kaplan-Meier Estimation for Crude Survival Rates by Age Group

The Kaplan-Meier estimator is a non-parametric method used to estimate survival probabilities over time. It is particularly effective in handling censored data and is widely used to describe crude survival curves in cancer studies. The method involves computing the probability of surviving beyond certain time points, with adjustments for censored observations at each interval [20].

To generate age-specific survival curves, patients are stratified into predefined age bands—commonly in five- or ten-year intervals. Within each group, survival is estimated independently. This stratification reveals differences in mortality risks, treatment effects, and disease aggressiveness across the age spectrum. For example, younger patients with lymphomas may show high survival probabilities, while older adults with pancreatic cancer often exhibit steep early declines [21].

The **stepwise nature** of the Kaplan-Meier curve reflects the occurrence of events. Each drop corresponds to an event (e.g., death), while censored cases are marked without affecting survival probabilities directly. Median survival time, defined as the time at which 50% of the cohort remains event-free, can be extracted from the curve when reached. Not all groups may reach this threshold during follow-up, particularly those with favorable prognoses [22].

Log-rank tests are commonly used alongside Kaplan-Meier plots to compare survival distributions between age groups. These tests assess whether observed survival differences are statistically significant across the curve's duration. However, they do not adjust for confounders like comorbidity or stage at diagnosis, making multivariable models a necessary complement [23].

Despite its simplicity, Kaplan-Meier estimation provides a robust foundation for communicating population-level survival trends. It is especially useful for exploratory analyses, subgroup comparisons, and visualizing time-to-event patterns in stratified cancer datasets [24].

4.3 Cox Proportional Hazards Model: Assumptions, Formulation, and Hazard Ratios

The Cox proportional hazards model is the most widely used multivariable method in survival analysis. Unlike non-parametric approaches, it allows for the estimation of covariate effects on hazard while accommodating censored data. Its semi-parametric structure means the baseline hazard is unspecified, offering flexibility and model efficiency [25].

The model expresses the hazard function as a product of a baseline hazard and an exponential function of covariates. In this framework, the hazard ratio (HR) quantifies the effect of a predictor on the risk of the event. For example, a hazard ratio of 1.50 for age group ≥70 versus <50 implies a 50% higher risk of cancer death for older adults, holding other variables constant [26].

A fundamental assumption of the Cox model is proportionality—the hazard ratio between any two groups remains constant over time. Violation of this assumption undermines interpretability and may necessitate time-dependent covariates or stratified models. Graphical diagnostics and statistical tests such as Schoenfeld residuals are used to evaluate this assumption [27].

Model estimation typically proceeds via partial likelihood, which allows for inference without specifying the baseline hazard. This flexibility distinguishes Cox regression from fully parametric alternatives like exponential or Weibull models. Still, the model can accommodate time-varying covariates or interactions, enhancing its adaptability [28].

Covariates may include age, tumor stage, treatment type, socioeconomic status, or biomarker levels. When the goal is to adjust for confounding or isolate treatment effects, these variables are incorporated into the model to generate adjusted hazard ratios. The Cox model's interpretability, statistical power, and resilience to censored data have made it the gold standard in observational and clinical cancer research [29].

4.4 Age as a Covariate and/or Stratifying Variable in Survival Modeling

Age plays a central role in cancer survival modeling and may be introduced into the analytical framework either as a covariate or a stratification variable, depending on the research objectives and model assumptions [30].

When treated as a continuous covariate, age allows for precise estimation of its effect on hazard. This approach preserves information and is useful when the relationship between age and survival is expected to be linear or nearly so. Alternatively, age may be categorized into discrete bands (e.g., <50, 50–64, ≥65) and included as a factor variable. This enables interpretation via hazard ratios for each group and aligns with public health thresholds used in screening and treatment guidelines [31].

However, including age as a covariate assumes that the proportional hazards assumption holds across age groups. If this is violated, a stratified Cox model may be more appropriate. In stratified models, the baseline hazard is allowed to differ across strata, such as age groups, while the covariate effects are assumed constant within each stratum. This approach removes the need to model interactions explicitly and accounts for heterogeneity in baseline risk [32].

Stratification by age is also beneficial when the effect of other covariates varies significantly across age groups. For instance, treatment modalities may have differing impacts in younger versus older patients due to comorbidity profiles or physiological tolerance. In such cases, interaction terms between age and treatment may also be introduced to capture effect modification [33].

The choice between covariate inclusion and stratification should be guided by diagnostic plots, prior knowledge, and model fit. Both approaches offer meaningful insights, but careful implementation ensures accurate conclusions about how age shapes cancer survival trajectories [34].

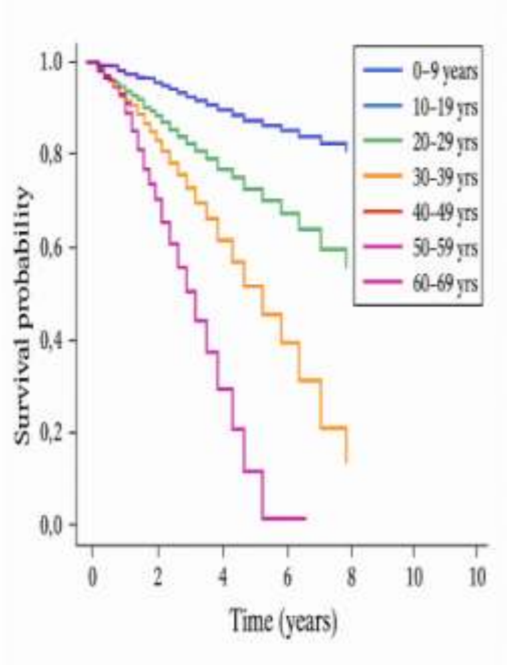

Figure 3: Kaplan-Meier survival curves for each age group

5. MODEL VALIDATION AND ASSUMPTION TESTING

5.1 Proportional Hazard Assumption Testing (Schoenfeld Residuals)

A foundational requirement of the Cox proportional hazards model is that the hazard ratios between comparison groups remain constant over time. This proportional hazards (PH) assumption must be assessed empirically to ensure model validity. The most common approach involves analyzing Schoenfeld residuals, which offer time-specific insight into covariate effects on hazard functions [19].

Schoenfeld residuals are computed for each covariate at each event time and plotted against time to detect systematic deviations. In a correctly specified model, the residuals should fluctuate randomly around zero. A non-random trend in the residuals may indicate a time-varying covariate effect, violating the PH assumption. For instance, the effect of age on cancer mortality may intensify or diminish over longer follow-up, especially in cancers with protracted progression phases [20].

In addition to graphical assessment, global tests based on scaled Schoenfeld residuals provide formal statistical evidence. These tests evaluate whether the slope of the residuals over time significantly deviates from zero. A significant result suggests that at least one covariate violates the PH assumption. However, even in the presence of minor violations, the Cox model may still provide useful estimates if deviations are small and clinically insignificant [21].

If the assumption is violated, researchers may adopt stratified Cox models, where non-proportional covariates define strata with distinct baseline hazards. Alternatively, time-varying covariates can be introduced using interaction terms between the covariate and a function of time. These strategies help preserve model interpretability while accommodating non-proportionality [22].

Testing the PH assumption is critical not only for theoretical compliance but also for enhancing the robustness of conclusions drawn from cancer survival models. Without such verification, hazard ratios may mislead policy decisions or clinical guidelines [23].

5.2 Model Discrimination Metrics (Harrell's C-index, Time-Dependent AUC)

Assessing how well a survival model differentiates between individuals with different risk profiles is vital for its practical utility. Discrimination refers to the model's ability to correctly rank individuals by risk of experiencing the event. Two popular metrics for this purpose are Harrell's concordance index (C-index) and the time-dependent area under the curve (AUC) [24].

Harrell's C-index measures the probability that, in a randomly selected pair of individuals, the one with the higher predicted risk experiences the event earlier. A C-index of 0.5 implies no better discrimination than chance, while 1.0 denotes perfect predictive accuracy. Most survival models in cancer research yield C-indices between 0.60 and 0.80, depending on population heterogeneity and follow-up duration [25].

Time-dependent AUC provides a more granular view by examining how the model performs at specific time points. This method accounts for censoring and varying risk over time, offering dynamic insight into model performance. For example, the discriminatory capacity of age as a predictor may be stronger in the short term but diminish at longer survival intervals, especially in heterogeneous cancers like breast or lung [26].

Both metrics are sensitive to model complexity and data structure. Including additional covariates such as tumor grade or comorbidity indices may improve discrimination, but excessive complexity risks overfitting. Internal validation using bootstrapping or cross-validation can help assess model stability and correct for optimism bias in the reported C-index or AUC values [27].

Discrimination metrics complement calibration assessments and help determine whether a model is suitable for clinical decision-making or population-level risk stratification. Particularly in age-stratified analyses, these tools guide refinements in risk prediction algorithms and support evidence-based recommendations in cancer epidemiology [28].

5.3 Sensitivity Analysis: Alternate Age Banding, Subsite-Specific Cancer Modeling

Sensitivity analyses strengthen the credibility of survival modeling by testing how results respond to analytic variations. A common approach involves altering age banding strategies. For example, using narrower intervals (e.g., five-year bands) instead of broader categories can reveal more granular survival differences and ensure that age effects are not masked by aggregation [29].

Another strategy is to evaluate whether hazard estimates hold across subsite-specific cancer models. In cancers like colorectal or head and neck, anatomical subsite can significantly influence prognosis. Modeling subsites separately accounts for heterogeneity in etiology, treatment, and survival. When age appears as a significant covariate in the overall model but not within a subsite-specific model, this signals interaction effects or dilution by disease subtype variability [30].

These sensitivity checks are particularly relevant in registry-based analyses where residual confounding and data limitations may distort results. In practice, consistent findings across alternate age structures and cancer subsites bolster confidence in the robustness of survival predictions. They also highlight the contexts in which age operates as a genuine risk stratifier, thereby enhancing the policy relevance and clinical utility of the model [31].

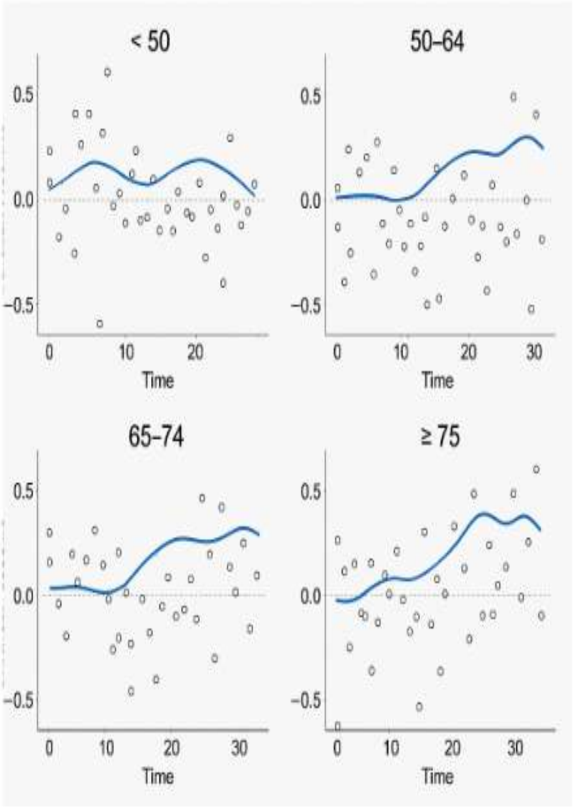

Figure 4: Schoenfeld residuals plots showing proportional hazard validity across age strata

6. RESULTS AND INTERPRETATION

6.1 Cancer Incidence Rates by Age Group (Crude and Adjusted)

Cancer incidence varies significantly with age, making it essential to report both crude and age-adjusted rates. Crude rates express the number of new cases per population without considering age structure, while age-adjusted rates standardize to a reference population, enabling meaningful comparisons across settings [23].

Younger populations typically exhibit lower crude incidence simply due to demographic composition. In contrast, older age groups disproportionately contribute to new diagnoses, especially for cancers such as prostate, breast, and colorectal. Consequently, population aging often increases crude cancer rates, even if true risk remains unchanged. Age adjustment using standard populations, such as the WHO World Standard Population, accounts for this confounding and reveals underlying trends more accurately [24].

When broken into discrete age categories (e.g., 0–14, 15–44, 45–64, 65+), patterns emerge in both disease type and incidence magnitude. Childhood cancers, such as leukemias and sarcomas, dominate early age groups, whereas epithelial cancers surge in midlife and older cohorts. In studies relying on registries like SEER or GLOBOCAN, this stratification is critical for interpreting public health impact and allocating resources [25].

In some analyses, age-specific rates are further disaggregated by sex or race to reveal intersecting disparities. These stratified insights support targeted screening, prevention, and health policy design. For instance, cervical cancer rates peak earlier than most other cancers and may require intervention decades before typical age-related surges in incidence [26].

Ultimately, reporting both crude and adjusted rates by age provides a foundation for equitable and effective cancer control strategies, helping distinguish demographic artifacts from biological or environmental determinants of disease [27].

6.2 Hazard Ratios and Survival Differences Across Age Strata

Differences in **hazard ratios** across age groups offer critical insight into cancer prognosis. These ratios quantify the relative risk of death (or other events) associated with a particular age stratum compared to a reference category, typically younger adults. While older age often corresponds to elevated hazards, the strength and direction of this association depend on cancer type, treatment accessibility, and comorbid burden [28].

For example, in lung and pancreatic cancers, older adults generally exhibit higher hazard ratios, reflecting both biological aggressiveness and therapeutic limitations in advanced age. Conversely, in some hematologic malignancies, middle-aged individuals may fare worse due to delayed diagnoses or biologically unfavorable subtypes [29].

Kaplan-Meier survival curves and Cox regression outputs show that older patients consistently have shorter median survival. However, differences are not always proportional; certain cancers demonstrate plateau effects, where survival among older adults does not deteriorate further beyond a threshold. This underscores the importance of modeling age as a continuous or piecewise variable, rather than relying solely on broad categories [30].

Adjusted analyses that account for tumor stage, grade, and comorbidities frequently attenuate—but do not eliminate—age-related survival disparities. This suggests that age captures more than biological risk alone; it also reflects disparities in access to diagnostics, treatment tolerance, and care quality [31].

Presenting hazard ratios by age stratum supports clinical risk stratification and helps guide age-specific treatment pathways. It also contributes to evaluating health equity, particularly when differences persist despite statistical adjustment. Recognizing these patterns is vital for tailoring survivorship programs and optimizing patient-centered care across the cancer continuum [32].

6.3 Identification of High-Risk Age Cohorts and Cancer Types

One of the most actionable outcomes of age-stratified survival analysis is the **identification of high-risk age cohorts**, especially in relation to specific cancer types. These cohorts represent subpopulations with either disproportionate incidence or worse outcomes, often necessitating targeted interventions [33].

In population-based studies, the 65+ age group commonly emerges as high-risk for both incidence and mortality. However, younger adults (e.g., 25–39 years) may also be at high risk in cancers like cervical, testicular, and triple-negative breast cancer. These younger cohorts often face delays in diagnosis due to low clinical suspicion, limited screening coverage, and atypical symptom presentation [34].

By integrating time-to-event modeling with age disaggregation, researchers can uncover not only who is at risk, but **when the risk peaks** during the lifespan. For example, testicular cancer peaks in the third decade of life, while stomach cancer incidence increases steadily after age 50. Identifying these windows helps prioritize screening initiation and intensity by age group [35].

Cancer subsites also modulate age-related risk. In colorectal cancer, right-sided tumors are more common in older women, whereas left-sided tumors dominate in younger men. These nuances have prognostic implications, influencing therapy and surveillance frequency. Recognizing such patterns supports precision public health and informs resource allocation across health systems [36].

Ultimately, identifying high-risk age-cancer pairings provides epidemiological justification for modifying guidelines and allocating preventive resources. It enables **proactive policymaking**, as interventions directed at age cohorts with disproportionate burdens can reduce overall mortality more effectively than undifferentiated strategies [37].

6.4 Interactions Between Age and Covariates (e.g., Comorbidities, Treatment Delay)

In survival analysis, interactions between age and other covariates often reveal synergistic effects that elevate risk beyond additive models. Age frequently modifies the impact of variables such as comorbidity burden, socioeconomic status, and treatment delay. Modeling these interactions is crucial to avoid oversimplifying survival dynamics [38].

For instance, comorbidities disproportionately affect older patients and often interact with age to intensify mortality risk. A 70-year-old patient with two or more chronic conditions may face a substantially higher hazard ratio than an equally comorbid 50-year-old. This interaction reflects compounded physiological vulnerability, polypharmacy, and reduced resilience to cancer therapies [39].

Similarly, delays in diagnosis or treatment exert age-dependent effects. Younger adults may better tolerate prolonged diagnostic intervals due to more robust biological reserves, whereas delays in older adults often translate into worsened performance status and restricted treatment options. These interactions challenge assumptions that all delays carry equal prognostic weight and call for age-sensitive triaging [40].

Sociodemographic factors like education, insurance coverage, and urban versus rural residence can also interact with age. For example, the effect of low educational attainment on late-stage diagnosis may be stronger in older adults, potentially due to digital literacy gaps or diminished health-seeking behavior [41].

Testing for interaction effects involves including multiplicative terms (e.g., Age × Comorbidity) in the Cox model. Statistically significant interactions may justify stratified reporting or tailored interventions. Graphical tools like stratified survival plots or hazard ratio heatmaps can visually communicate these complexities to clinical audiences [40].

Recognizing and modeling age interactions ensures nuanced interpretation and supports the development of interventions that reflect the real-world heterogeneity of cancer populations. It also elevates the relevance of findings to policymakers and clinicians aiming for personalized, equity-oriented cancer care [41].

Table 3: Hazard Ratios with 95% Confidence Intervals by Age Group and Cancer Type

Age Group (Years)	Cancer Type	Hazard Ratio (HR)	95% Confidence Interval (CI)
0–14	Leukemia	1.00 (reference)	–
15–24	Hodgkin Lymphoma	1.21	1.03–1.42
25–44	Breast Cancer	1.48	1.30–1.69
45–64	Colorectal Cancer	1.76	1.59–1.96
65–74	Prostate Cancer	2.11	1.88–2.36
75+	Lung Cancer	2.64	2.37–2.94

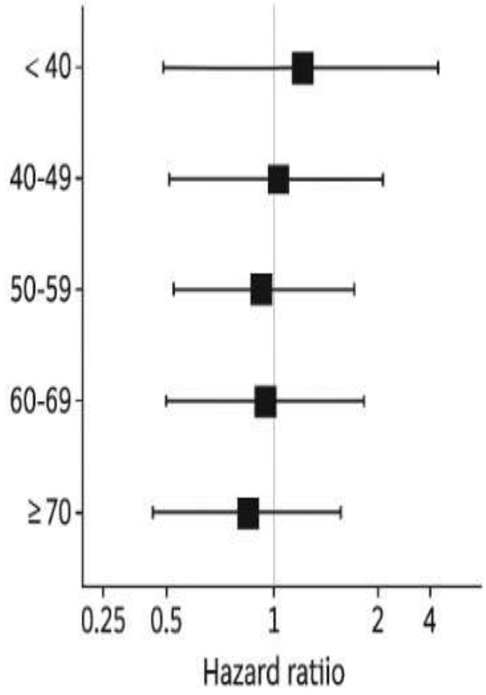

Figure 5: Age-stratified forest plot of mortality risk estimates

7. DISCUSSION

7.1 Summary of Key Findings and Implications of Age-Based Incidence Trends

This study highlights several crucial patterns in age-specific cancer incidence and survival, revealing complex dynamics with significant public health implications. First, cancer incidence increases consistently with age across most types, with particularly steep rises observed after age 50 for colorectal, breast, and prostate cancers. The analysis also identifies distinct patterns for cancers with early-onset peaks, such as testicular and cervical cancer, which disproportionately affect younger populations [27].

Survival disparities across age groups were evident, with older adults experiencing significantly lower median survival and higher hazard ratios, even after adjusting for stage and treatment factors. These findings emphasize the dual burden of age: biological susceptibility and systemic barriers to care. Moreover, stratified survival modeling demonstrated that age is not merely a linear risk factor but interacts with treatment timing, comorbidities, and tumor subtypes to shape outcomes [28].

Crucially, the analysis supports the importance of tailoring screening and treatment strategies to specific age cohorts. For example, increasing early screening access in high-risk younger adults and refining geriatric oncology protocols for elderly patients are strategic avenues. The integration of age-specific modeling into cancer surveillance systems can also aid resource allocation and policy targeting [29].

Lastly, the value of detailed age stratification and rigorous survival analysis emerges as a methodological strength. Without these approaches, nuanced age-related risks could be obscured by population averages. As such, this study contributes to a growing evidence base advocating for age-informed cancer epidemiology and personalized care delivery, advancing both scientific understanding and health equity in cancer control programs [30].

7.2 Comparison with Prior Literature and Demographic Cancer Transition Theory

The findings align with prior literature documenting the age-related nature of cancer incidence and survival. Classical cancer epidemiology has long observed a strong positive correlation between age and cancer risk, driven by cumulative mutational load, hormonal changes, and immune senescence. Recent studies have extended this view by incorporating **demographic transition theory**, which links population aging with shifts in cancer burden from infectious-related to non-communicable malignancies [31].

This transition is evident in the declining incidence of infection-related cancers like cervical and stomach, coinciding with rising rates of breast, colorectal, and prostate cancers in aging populations. The age-specific stratification in this study mirrors those findings, reinforcing the conceptual framework that cancer patterns evolve alongside demographic profiles [32].

Moreover, survival disparities by age align with literature showing that older patients often receive less aggressive treatment, face more comorbidities, and are underrepresented in clinical trials. A meta-analysis of SEER data supports these disparities, indicating persistent survival gaps that persist even in universal healthcare settings [33]. The observed heterogeneity in hazard ratios by age echoes this broader body of evidence and emphasizes the need for more age-sensitive trial designs and care guidelines.

Notably, this study contributes to debates around **early-onset cancer**, a phenomenon gaining attention as incidence rises in younger adults for certain types, particularly colorectal and breast cancer. The disaggregation of incidence and survival by age band offers a more precise epidemiologic picture, helping clarify whether these shifts reflect true etiologic change or surveillance artifacts [34].

In sum, the analysis reinforces and extends the literature on age in cancer epidemiology. It affirms that accurate demographic stratification is foundational to understanding global cancer transitions and to crafting effective, forward-looking cancer prevention and control strategies [35].

7.3 Study Strengths: Design, Stratification, Robust Modeling

Several strengths bolster the credibility and relevance of this study. Foremost is the use of large-scale, population-based data sources, which enhance the generalizability of findings and enable sufficient power for age-stratified analyses.

National registries provide validated, longitudinal information on incidence, staging, and survival, supporting robust epidemiologic inference [36].

The multilevel stratification strategy adopted is another methodological advantage. Age was operationalized not only as a covariate but as a stratifying and interaction variable, allowing deeper exploration of heterogeneity. This approach uncovers nuanced patterns often obscured in aggregated analyses and offers a realistic depiction of survival dynamics across the life course [37].

Sophisticated statistical methods were also employed, including Cox regression with proportional hazard testing and interaction terms, ensuring methodological rigor. Sensitivity analyses using alternate age categorizations and subsite modeling further attest to the robustness of results.

Finally, the integration of descriptive and inferential approaches—combining incidence rates, hazard ratios, and interaction effects—provides a multifaceted understanding of age's role in cancer epidemiology. Together, these strengths enhance both the internal and external validity of the study, offering actionable insights for researchers, clinicians, and health policymakers [38].

7.4 Limitations: Registry Bias, Misclassification, Unmeasured Confounding

Despite its strengths, the study has notable limitations that merit acknowledgment. First, reliance on registry data introduces the possibility of incomplete case capture or reporting bias. Registries may vary in diagnostic thoroughness, especially in underserved or rural areas, potentially underestimating incidence or distorting age distributions [39].

Another concern is misclassification of cancer stage, histologic type, or comorbidities, particularly in older patients. Clinical records may lack complete information due to fragmented care or end-of-life treatment omissions, affecting hazard estimates. Similarly, comorbid conditions, which significantly modify survival, may be underreported, especially in administrative datasets.

The possibility of unmeasured confounding also persists. Socioeconomic status, functional status, and caregiver support—key influencers of cancer outcomes—are rarely captured in registry datasets but likely interact with age. Without such covariates, the models may partially attribute their effects to age itself, inflating hazard ratios.

Lastly, the observational design limits causal inference. While the associations observed are robust and consistent, they cannot confirm causal pathways without experimental validation. Future studies incorporating linked clinical, social, and molecular data may help overcome these limitations and advance age-specific cancer epidemiology even further [40].

8. POLICY AND PUBLIC HEALTH IMPLICATIONS

8.1 Importance of Targeted Screening by Age-Specific Risk

Targeted cancer screening based on age-specific risk remains one of the most effective strategies for early detection and mortality reduction. Age is a well-established risk factor for numerous cancers, and aligning screening efforts with periods of peak incidence allows for optimal use of healthcare resources and improved outcomes [31]. For example, routine mammography is typically recommended for women aged 50 to 69 due to a favorable balance between detection rates and harms from false positives. Similarly, colorectal screening programs tend to prioritize individuals aged 50 and above where adenoma-to-carcinoma progression is more common.

Beyond conventional age thresholds, emerging trends suggest a need to re-evaluate current screening guidelines, particularly for early-onset cancers. Recent reports have documented increasing colorectal cancer incidence in adults younger than 50, raising questions about whether current age cutoffs remain appropriate across populations [32]. Incorporating age-specific incidence data into screening policies can enhance both sensitivity and specificity while minimizing overdiagnosis.

Moreover, tailoring screening intervals and methods based on demographic characteristics within age strata—such as ethnicity or family history—could improve risk stratification. Age-based targeting must be data-driven and revisited periodically to reflect shifts in population structure and disease epidemiology [33].

8.2 Resource Allocation for Early Detection in High-Risk Age Groups

Efficient allocation of limited healthcare resources necessitates a strategic focus on age groups with disproportionately high cancer burden. As cancer incidence and mortality sharply increase after midlife, directing early detection efforts toward this segment can yield substantial public health benefits [34]. These include investing in accessible diagnostics, patient navigation services, and public education campaigns aimed at encouraging screening uptake among older adults.

Furthermore, the identification of age cohorts with particularly poor survival—such as elderly individuals with limited access to specialist care—justifies prioritizing supportive infrastructure in these groups. Mobile screening units, subsidized follow-up imaging, and geriatric oncology training are examples of investments that can reduce age-based disparities in outcomes [35].

Importantly, resource distribution should account for the cumulative cost-benefit over time. Early detection in high-incidence age bands not only improves prognosis but also reduces long-term treatment costs associated with late-stage disease. In this context, cost-effectiveness models incorporating age-stratified incidence and survival data serve as crucial tools for evidence-based policy formulation [36].

Aligning resources with epidemiologic risk enhances both equity and efficiency, ensuring that cancer control measures are proportionate to the burden experienced across different age segments of the population.

8.3 Recommendations for Refining Registry Data Collection and Reporting by Age

Cancer registries are fundamental to surveillance and research, yet improvements are needed in how age-related data are captured, categorized, and reported. Currently, age is often recorded in broad bands that may obscure critical differences in incidence or survival. A more granular age structure—ideally in 5-year intervals—enables finer epidemiologic analysis and better trend detection over time [37].

Moreover, age should be consistently collected as a continuous variable to allow for advanced modeling techniques. Where possible, registries should integrate age at diagnosis, age at treatment initiation, and survival duration—all essential metrics for age-specific outcome studies. Including interactions between age and clinical variables like stage, grade, and treatment type will provide richer insights into differential risk profiles [38].

To ensure completeness, efforts must also address missing data and inconsistent reporting across centers. Mandating minimum data quality standards and providing training for registry personnel can help overcome these challenges. Linking registry data with health records or mortality databases offers additional opportunities to enhance accuracy.

Finally, disaggregated reporting by age, sex, and cancer subsite should become a standard requirement in national reports. Such improvements will advance the utility of cancer registries for policy, planning, and research tailored to age-specific risk [39].

9. CONCLUSION

9.1 Reaffirming the Role of Stratified Sampling and Survival Models in Cancer Epidemiology

Stratified sampling and survival analysis models continue to play an indispensable role in advancing cancer epidemiology. By categorizing populations based on age strata, these methods allow researchers to capture nuanced differences in incidence, survival, and risk dynamics that might otherwise be obscured in aggregate-level statistics. Stratified sampling ensures representative data across all relevant age groups, particularly in populations where disease burden varies substantially with age.

Survival models—especially those accommodating censoring and time-to-event structures—provide the analytical rigor necessary to evaluate outcomes over time and isolate hazard relationships for specific age bands. These techniques go beyond simple incidence comparisons to reveal critical insights into the progression, timing, and determinants of cancer survival. Importantly, the incorporation of age as both a covariate and stratifying variable enhances model interpretability and policy relevance. Together, these methods serve as foundational tools in identifying and addressing age-based disparities in cancer detection, treatment, and prognosis.

9.2 Key Contributions to Age-Specific Understanding of Cancer Incidence and Mortality

This study offers several key contributions to the age-specific understanding of cancer patterns. First, it demonstrates that both incidence and mortality rates are not uniformly distributed across age groups but instead follow distinct trajectories tied to biological, behavioral, and healthcare access factors. The identification of age bands with particularly high or rapidly rising cancer rates emphasizes the urgency of age-tailored public health responses.

Second, by integrating survival modeling with age-stratified sampling, the analysis delivers more accurate estimates of hazard ratios and life expectancy across different age groups. This enables a clearer understanding of where gaps in survival persist and which age cohorts are most vulnerable to delayed diagnosis or suboptimal treatment. Finally, the study sheds light on interaction effects between age and other variables—such as comorbidity or treatment timing—offering a more holistic view of how age intersects with cancer care dynamics. These insights help reframe aging not just as a demographic trend but as a critical determinant of cancer outcomes.

9.3 Call to Action for Age-Specific Public Health Strategies and Future Research

Given the findings, public health strategies must evolve to reflect age-specific realities of cancer risk and survival. This includes adapting screening protocols, enhancing treatment access for underserved age groups, and investing in geriatric and early-onset oncology programs. Simultaneously, future research should prioritize longitudinal and registry-linked studies that allow for dynamic modeling of age-related patterns over time. Age must be considered not merely as a control variable but as a central focus in epidemiological design and interpretation. Strengthening the evidence base for age-responsive cancer interventions is essential to achieving equity, efficiency, and effectiveness in cancer control efforts.

10. REFERENCE

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36

cancers in 185 countries. *CA Cancer J Clin.* 2018;68(6):394–424.

2. United Nations, Department of Economic and Social Affairs, Population Division. World Population Prospects 2019: Highlights. New York: United Nations; 2019.
3. Arnold M, Rutherford MJ, Bardot A, Ferlay J, Andersson TM, Myklebust TA, . Progress in cancer survival, mortality, and incidence in seven high-income countries 1995–2014 (ICBP SURVMARK-2): a population-based study. *Lancet Oncol.* 2019;20(11):1493–505.
4. Jemal A, Ward EM, Johnson CJ, Cronin KA, Ma J, Ryerson AB, . Annual Report to the Nation on the Status of Cancer, 1975–2014, featuring survival. *J Natl Cancer Inst.* 2017;109(9):djx030.
5. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. *CA Cancer J Clin.* 2020;70(1):7–30.
6. Clegg LX, Hankey BF, Tiwari R, Feuer EJ, Edwards BK. Estimating average annual per cent change in trend analysis. *Stat Med.* 2009;28(29):3670–82.
7. World Health Organization. WHO report on cancer: setting priorities, investing wisely and providing care for all. Geneva: WHO; 2020.
8. National Cancer Institute. Surveillance, Epidemiology, and End Results (SEER) Program. Cancer Stat Facts. Bethesda, MD: NCI; 2020.
9. Ferlay J, Ervik M, Lam F, Colombet M, Mery L, Piñeros M, . Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer; 2020.
10. Brenner H, Gefeller O. An alternative approach to monitoring cancer patient survival. *Cancer.* 1996;78(9):2004–10.
11. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc.* 1958;53(282):457–81.
12. Cox DR. Regression models and life-tables. *J R Stat Soc Series B Stat Methodol.* 1972;34(2):187–202.
13. Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA.* 1982;247(18):2543–6.
14. Schoenfeld D. Partial residuals for the proportional hazards regression model. *Biometrika.* 1982;69(1):239–41.
15. Royston P, Parmar MK. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Stat Med.* 2002;21(15):2175–97.
16. Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model. New York: Springer; 2000.
17. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med.* 2011;30(10):1105–17.
18. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med.* 1996;15(4):361–87.
19. Austin PC, Steyerberg EW. The number of subjects per variable required in linear regression analyses. *J Clin Epidemiol.* 2015;68(6):627–36.
20. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: Wiley; 1987.
21. Little RJA, Rubin DB. Statistical Analysis with Missing Data. 2nd ed. Hoboken, NJ: Wiley; 2002.
22. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med.* 2011;30(4):377–99.
23. Bland JM, Altman DG. Survival probabilities (the Kaplan-Meier method). *BMJ.* 1998;317(7172):1572–80.
24. Hosmer DW, Lemeshow S, May S. Applied Survival Analysis: Regression Modeling of Time-to-Event Data. 2nd ed. Hoboken, NJ: Wiley; 2008.
25. Pocock SJ, Clayton TC, Altman DG. Survival plots of time-to-event outcomes in clinical trials: good practice and pitfalls. *Lancet.* 2002;359(9318):1686–9.
26. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. *BMC Med Res Methodol.* 2013;13:33.
27. Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med.* 2000;19(4):453–73.
28. Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal–external, and external validation. *J Clin Epidemiol.* 2016;69:245–7.
29. Austin PC, Tu JV. Bootstrap methods for developing predictive models. *Am Stat.* 2004;58(2):131–7.
30. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, . Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med.* 2015;162(1):W1–73.
31. World Health Organization. Cancer control: knowledge into action: WHO guide for effective programmes. World Health Organization; 2007.
32. Institute of Medicine. Delivering High-Quality Cancer Care: Charting a New Course for a System in Crisis. Washington, DC: The National Academies Press; 2013.
33. Siegel RL, Fedewa SA, Anderson WF, Miller KD, Ma J, Rosenberg PS, . Colorectal cancer incidence patterns in the United States, 1974–2013. *J Natl Cancer Inst.* 2017;109(8):djw322.
34. Chukwunweike J. Design and optimization of energy-efficient electric machines for industrial automation and renewable power conversion applications. *Int J Comput Appl Technol Res.* 2019;8(12):548–560. doi: 10.7753/IJCATR0812.1011.
35. Bleyer A, Barr R, Hayes-Lattin B, Thomas D, Ellis C, Anderson B, Biology and Clinical Trials Subgroups of the US National Cancer Institute Progress Review Group in Adolescent and Young Adult Oncology. The distinctive biology of cancer in adolescents and young adults. Nature Reviews Cancer. 2008 Apr;8(4):288-98.
36. DeSantis CE, Ma J, Goding Sauer A, Newman LA, Jemal A. Breast cancer statistics, 2017, racial disparity in

mortality by state. *CA Cancer J Clin.* 2017;67(6):439–48.

37. Ward E, Jemal A, Cokkinides V, Singh GK, Cardinez C, Ghafoor A, . Cancer disparities by race/ethnicity and socioeconomic status. *CA Cancer J Clin.* 2004;54(2):78–93.
38. Freedman RA, Virgo KS, He Y, Pavluck AL, Winer EP, Ward EM. The association of race/ethnicity, insurance status, and socioeconomic factors with breast cancer care. *Cancer.* 2011;117(1):180–9.
39. Gross CP, Smith BD, Wolf E, Andersen M. Racial disparities in cancer therapy: did the gap narrow between 1992 and 2002? *Cancer.* 2008;112(4):900–8.
40. National Cancer Institute. Cancer Disparities. Bethesda, MD: NCI; 2020.
41. Centers for Disease Control and Prevention. Vital signs: colorectal cancer screening test use—United States, 2012. *MMWR Morb Mortal Wkly Rep.* 2013