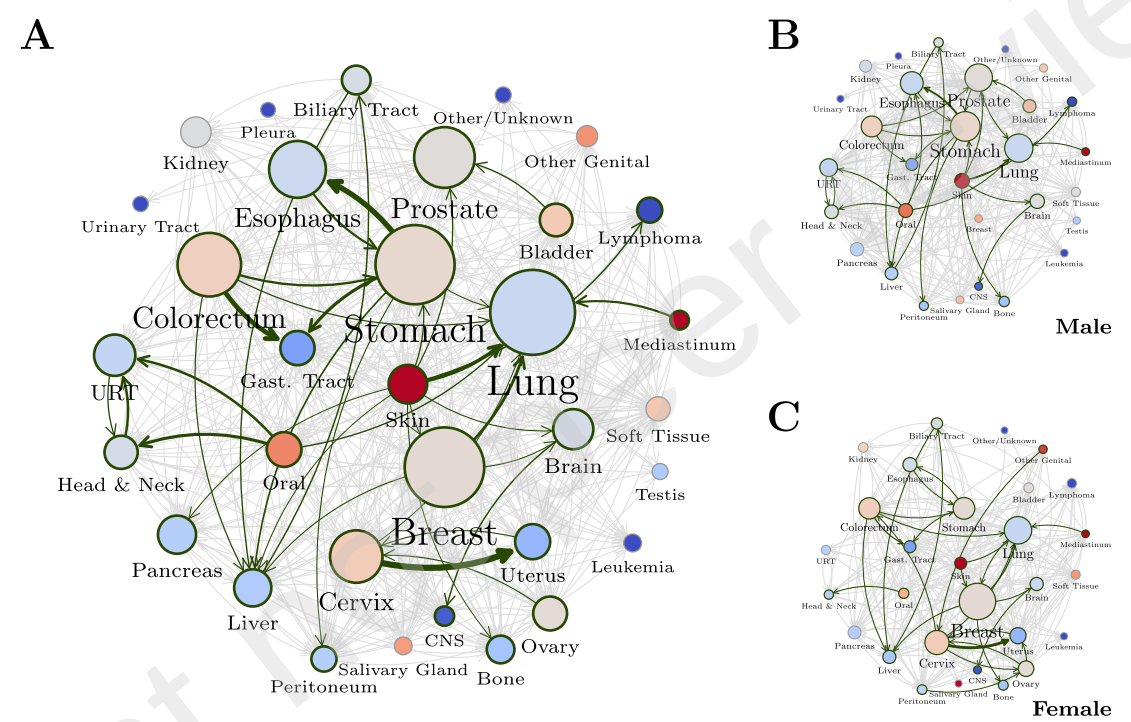

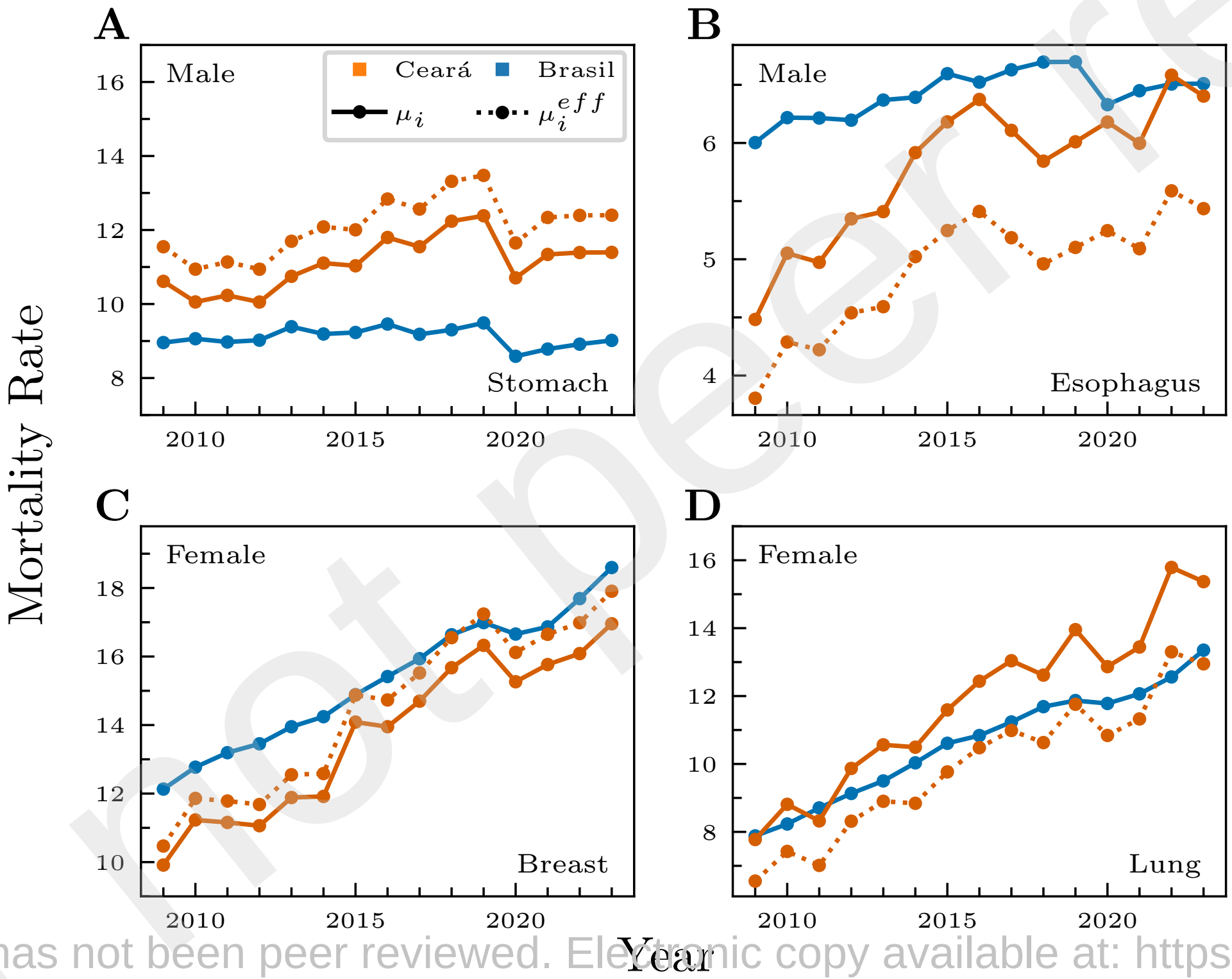

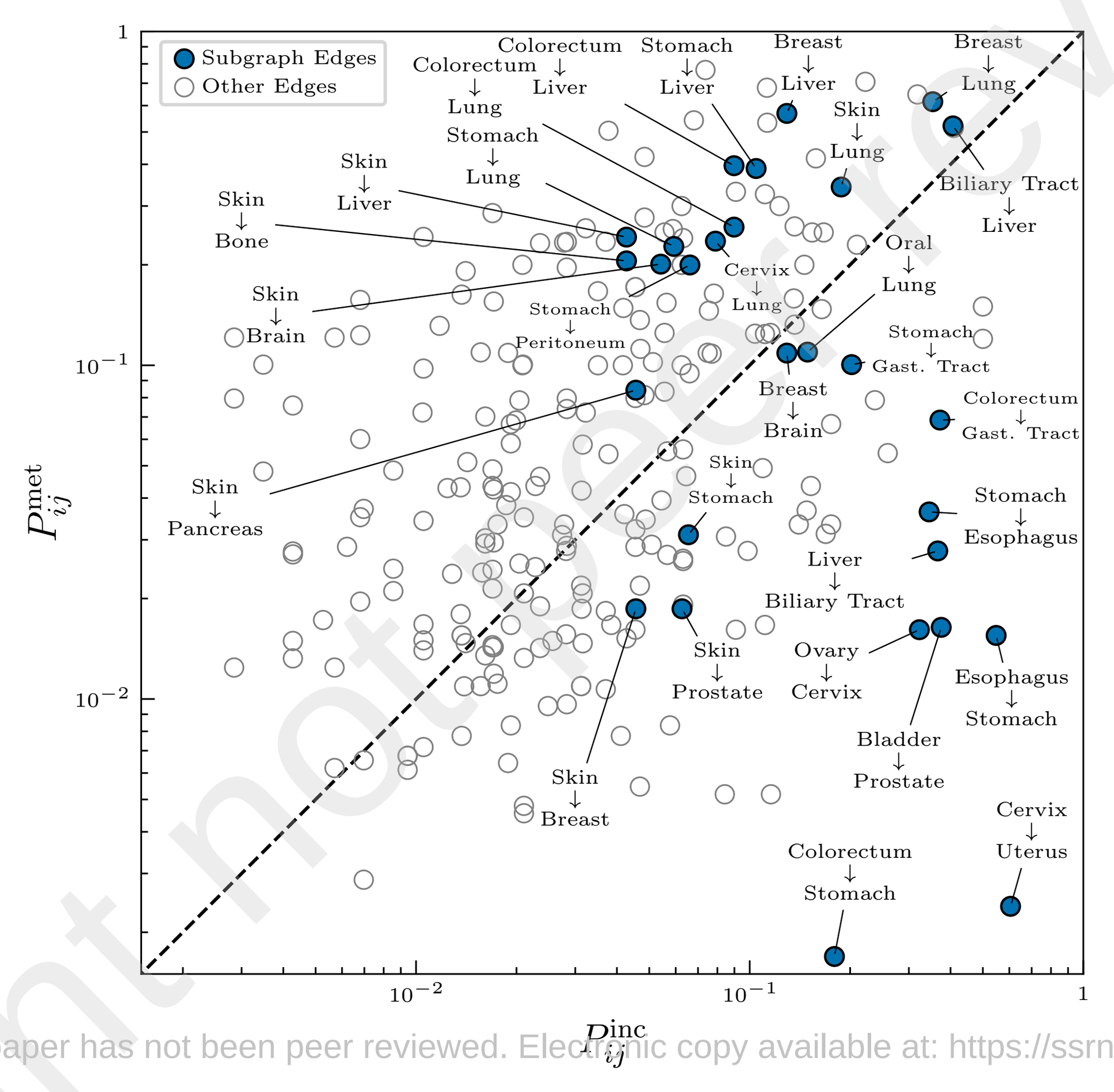

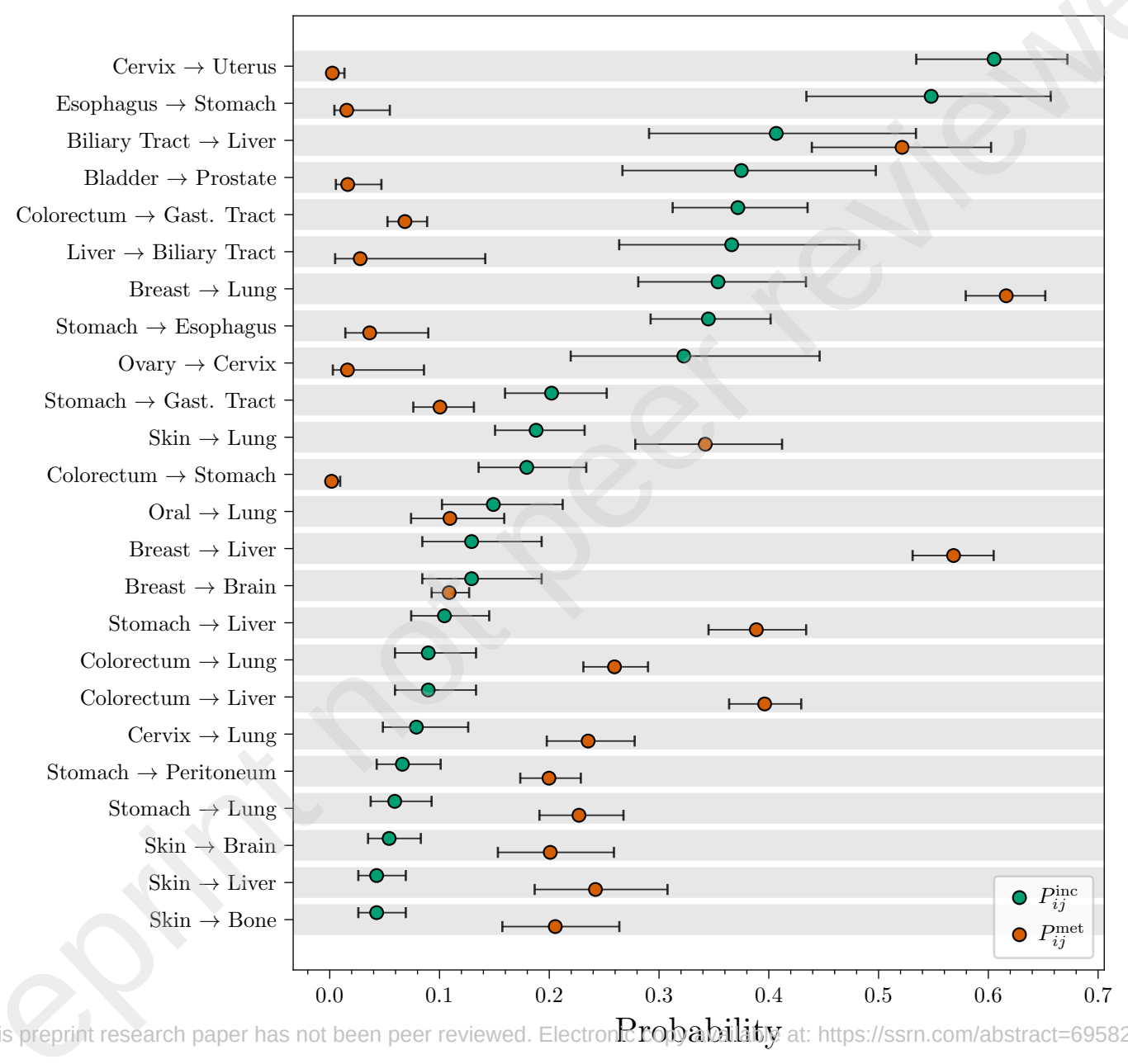

ICD-10 Code	Description (EN)
C00	Malignant neoplasm of lip
C01	Malignant neoplasm of base of tongue
C02	Malignant neoplasm of other/unspecified parts of tongue
C03	Malignant neoplasm of gum
C04	Malignant neoplasm of floor of mouth
C05	Malignant neoplasm of palate
C06	Malignant neoplasm of other/unspecified parts of mouth
C07	Malignant neoplasm of parotid gland
C08	Malignant neoplasm of other/unspecified major salivary glands
C09	Malignant neoplasm of tonsil
C10	Malignant neoplasm of oropharynx
C11	Malignant neoplasm of nasopharynx
C12	Malignant neoplasm of pyriform sinus
C13	Malignant neoplasm of hypopharynx
C14	Malignant neoplasm of other/ill-defined sites in lip, oral cavity and pharynx
C15	Malignant neoplasm of esophagus
C16	Malignant neoplasm of stomach
C17	Malignant neoplasm of small intestine
C18	Malignant neoplasm of colon
C19	Malignant neoplasm of rectosigmoid junction
C20	Malignant neoplasm of rectum
C21	Malignant neoplasm of anus and anal canal
C22	Malignant neoplasm of liver and intrahepatic bile ducts
C23	Malignant neoplasm of gallbladder
C24	Malignant neoplasm of other/unspecified parts of biliary tract
C25	Malignant neoplasm of pancreas
C26	Malignant neoplasm of other/ill-defined digestive organs
C30	Malignant neoplasm of nasal cavity and middle ear
C31	Malignant neoplasm of accessory sinuses
C32	Malignant neoplasm of larynx
C33	Malignant neoplasm of trachea
C34	Malignant neoplasm of bronchus and lung
C37	Malignant neoplasm of thymus
C38	Malignant neoplasm of heart, mediastinum and pleura
C39	Malignant neoplasm of other/ill-defined sites in respiratory system
C40	Malignant neoplasm of bone and articular cartilage of limbs
C41	Malignant neoplasm of bone and articular cartilage of other/unspecified sites
C43	Malignant melanoma of skin
C44	Other malignant neoplasms of skin
C45	Mesothelioma
C46	Kaposi's sarcoma
C47	Malignant neoplasm of peripheral nerves and autonomic nervous system
C48	Malignant neoplasm of retroperitoneum and peritoneum

C49	Malignant neoplasm of other connective and soft tissue
C50	Malignant neoplasm of breast
C51	Malignant neoplasm of vulva
C52	Malignant neoplasm of vagina
C53	Malignant neoplasm of cervix uteri
C54	Malignant neoplasm of corpus uteri
C55	Malignant neoplasm of uterus, part unspecified
C56	Malignant neoplasm of ovary
C57	Malignant neoplasm of other/unspecified female genital organs
C58	Malignant neoplasm of placenta
C60	Malignant neoplasm of penis
C61	Malignant neoplasm of prostate
C62	Malignant neoplasm of testis
C63	Malignant neoplasm of other/unspecified male genital organs
C64	Malignant neoplasm of kidney, except renal pelvis
C65	Malignant neoplasm of renal pelvis
C66	Malignant neoplasm of ureter
C67	Malignant neoplasm of bladder
C68	Malignant neoplasm of other/unspecified urinary organs
C69	Malignant neoplasm of eye and adnexa
C70	Malignant neoplasm of meninges
C71	Malignant neoplasm of brain
C72	Malignant neoplasm of spinal cord, cranial nerves and other parts of CNS
C73	Malignant neoplasm of thyroid gland
C74	Malignant neoplasm of adrenal gland
C75	Malignant neoplasm of other endocrine glands and related structures
C76	Malignant neoplasm of other/ill-defined sites
C77	Secondary malignant neoplasm of lymph nodes
C78	Secondary malignant neoplasm of respiratory and digestive organs
C79	Secondary malignant neoplasm of other and unspecified sites
C80	Malignant neoplasm without specification of site
C81	Hodgkin's lymphoma
C82	Follicular lymphoma
C83	Non-follicular lymphoma
C84	Mature T/NK-cell lymphomas
C85	Other and unspecified types of non-Hodgkin lymphoma
C86	Other specified types of T/NK-cell lymphoma
C88	Malignant immunoproliferative diseases
C90	Multiple myeloma and malignant plasma cell neoplasms
C91	Lymphoid leukaemia
C92	Myeloid leukaemia
C93	Monocytic leukaemia
C94	Other leukaemias of specified cell type
C95	Leukaemia of unspecified cell type

C96	Other and unspecified malignant neoplasms of lymphoid, haematopoietic tissue
C97	Malignant neoplasms of independent (primary) multiple sites
D00	Carcinoma in situ of oral cavity, oesophagus and stomach
D01	Carcinoma in situ of other and unspecified digestive organs
D02	Carcinoma in situ of middle ear and respiratory system
D03	Melanoma in situ
D04	Carcinoma in situ of skin
D05	Carcinoma in situ of breast
D06	Carcinoma in situ of cervix uteri
D07	Carcinoma in situ of other and unspecified genital organs
D09	Carcinoma in situ of other and unspecified sites

Description (PT)	Broad Organ Class
Neoplasia maligna do lábio	oral cavity
Neoplasia maligna da base da língua	oral cavity
Neoplasia maligna de outras partes e de partes não especificadas da língua	oral cavity
Neoplasia maligna da gengiva	oral cavity
Neoplasia maligna do assoalho da boca	oral cavity
Neoplasia maligna do palato	oral cavity
Neoplasia maligna de outras partes e de partes não especificadas da boca	oral cavity
Neoplasia maligna da glândula parótida	salivary gland
Neoplasia maligna de outras glândulas salivares maiores e das não especificadas	salivary gland
Neoplasia maligna da tonsila	head and neck
Neoplasia maligna da orofaringe	head and neck
Neoplasia maligna da nasofaringe	head and neck
Neoplasia maligna do seio piriforme	head and neck
Neoplasia maligna da hipofaringe	head and neck
Neoplasia maligna de outras localizações e de localizações mal definidas do lábio, cavidade oral e faringe	head and neck
Neoplasia maligna do esôfago	esophagus
Neoplasia maligna do estômago	stomach
Neoplasia maligna do intestino delgado	gastrointestinal tract
Neoplasia maligna do cólon	colorectum
Neoplasia maligna da junção retossigmóide	colorectum
Neoplasia maligna do reto	colorectum
Neoplasia maligna do ânus e do canal anal	colorectum
Neoplasia maligna do fígado e das vias biliares intra-hepáticas	liver
Neoplasia maligna da vesícula biliar	biliary tract
Neoplasia maligna de outras partes e de partes não especificadas das vias biliares	biliary tract
Neoplasia maligna do pâncreas	pancreas
Neoplasia maligna de outros órgãos digestivos e de localizações mal definidas	gastrointestinal tract
Neoplasia maligna da cavidade nasal e do ouvido médio	head and neck
Neoplasia maligna dos seios da face	head and neck
Neoplasia maligna da laringe	upper respiratory tract
Neoplasia maligna da traquéia	upper respiratory tract
Neoplasia maligna dos brônquios e dos pulmões	lung
Neoplasia maligna do timo	mediastinum
Neoplasia maligna do coração, mediastino e pleura	mediastinum
Neoplasia maligna de outras localizações e de localizações mal definidas do sistema respiratório	lung
Neoplasia maligna dos ossos e cartilagens articulares dos membros	bone
Neoplasia maligna dos ossos e cartilagens articulares de outras localizações e de localizações não especificadas	bone
Melanoma maligno da pele	skin
Outras neoplasias malignas da pele	skin
Mesotelioma	pleura
Sarcoma de Kaposi	soft tissue
Neoplasia maligna dos nervos periféricos e do sistema nervoso autônomo	soft tissue
Neoplasia maligna do retroperitônio e do peritônio	peritoneum

Neoplasia maligna do tecido conjuntivo e de outros tecidos moles	soft tissue
Neoplasia maligna da mama	breast
Neoplasia maligna da vulva	other genital
Neoplasia maligna da vagina	other genital
Neoplasia maligna do colo do útero	cervix
Neoplasia maligna do corpo do útero	uterus
Neoplasia maligna do útero, porção não especificada	uterus
Neoplasia maligna do ovário	ovary
Neoplasia maligna de outros órgãos genitais femininos e dos não especificados	other genital
Neoplasia maligna da placenta	other genital
Neoplasia maligna do pênis	other genital
Neoplasia maligna da próstata	prostate
Neoplasia maligna dos testículos	testis
Neoplasia maligna de outros órgãos genitais masculinos e dos não especificados	other genital
Neoplasia maligna do rim, exceto pelve renal	kidney
Neoplasia maligna da pelve renal	kidney
Neoplasia maligna do ureter	urinary tract
Neoplasia maligna da bexiga	bladder
Neoplasia maligna de outros órgãos urinários e dos não especificados	urinary tract
Neoplasia maligna do olho e anexos	other/unknown
Neoplasia maligna das meninges	brain
Neoplasia maligna do encéfalo	brain
Neoplasia maligna da medula espinal, dos nervos cranianos e de outras partes do SNC	central nervous system
Neoplasia maligna da glândula tireoide	thyroid
Neoplasia maligna da glândula suprarrenal	adrenal gland
Neoplasia maligna de outras glândulas endócrinas e estruturas relacionadas	neuroendocrine
Neoplasia maligna de outras localizações e de localizações mal definidas	other/unknown
Neoplasia maligna secundária dos gânglios linfáticos	lymph node
Neoplasia maligna secundária dos órgãos respiratórios e digestivos	other/unknown
Neoplasia maligna secundária de outras localizações e das não especificadas	other/unknown
Neoplasia maligna sem especificação de localização	unknown primary
Doença de Hodgkin	lymphoma
Linfoma folicular	lymphoma
Linfoma não folicular	lymphoma
Linfomas de células T/NK maduras	lymphoma
Linfoma não-Hodgkin de outros tipos e de tipo não especificado	lymphoma
Outros tipos especificados de linfoma de células T/NK	lymphoma
Doenças imunoproliferativas malignas	lymphoma
Mieloma múltiplo e neoplasias malignas de plasmócitos	bone
Leucemia linfoide	leukemia
Leucemia mieloide	leukemia
Leucemia monocítica	leukemia
Outras leucemias de células de tipo especificado	leukemia
Leucemia de tipo celular não especificado	leukemia

Outras neoplasias malignas e as não especificadas dos tecidos linfoide, hematopoético e relacionados	lymphoma
Neoplasias malignas de localizações múltiplas independentes (primárias)	other/unknown
Carcinoma in situ da cavidade oral, esôfago e estômago	upper gastrointestinal
Carcinoma in situ de outros órgãos digestivos e dos não especificados	gastrointestinal tract
Carcinoma in situ do ouvido médio e do sistema respiratório	lung
Melanoma in situ	skin
Carcinoma in situ da pele	skin
Carcinoma in situ da mama	breast
Carcinoma in situ do colo do útero	cervix
Carcinoma in situ de outros órgãos genitais e dos não especificados	other genital
Carcinoma in situ de outras localizações e das não especificadas	other/unknown

Original Term	Broad Organ Class
abdom nodes	lymph node
abdominal LN	lymph node
adrenal glands	adrenal gland
adrenal_gland	adrenal gland
andomen cav	peritoneum
anus	colorectum
appendix	colorectum
auxiliary nodes	lymph node
axillary LN	lymph node
bile duct	biliary tract
biliary_cancer	biliary tract
bladder	bladder
bone	bone
bone_marrow	bone
brain cerebral	brain
brain dural	brain
brain	brain
branchial cyst	head and neck
breast	breast
breast_cancer	breast
cervical nodes	lymph node
cervical_cancer	cervix
cervix	cervix
colon	colorectum
colorectal_cancer	colorectum
diaphragm	mediastinum
duodenum	gastrointestinal tract
esophageal_and_gastric_cancer	upper gastrointestinal
esophagus	esophagus
eye	other/unknown
fallopian tubes	other genital
gallbladder	biliary tract
gastroint tract	gastrointestinal tract
head_and_neck_cancer	head and neck
heart	mediastinum
inguinal nodes	lymph node
intestinal LN	lymph node
kidney	kidney
kidney_cancer	kidney
large intestine	colorectum
larynx	upper respiratory tract
limph nodes dist	lymph node
limph nodes reg	lymph node

lip	oral cavity
liver	liver
liver_cancer	liver
lung	lung
lung_cancer	lung
mediast nodes	lymph node
mediastinal LN	lymph node
melanoma	skin
meninges	brain
mesentery	peritoneum
neuroendocrine_cancer	neuroendocrine
non_regional_lymph_nodes	lymph node
omentum	peritoneum
other	other/unknown
other_site	other/unknown
ovarian_cancer	ovary
ovaries	ovary
ovary	ovary
pancreas	pancreas
pancreatic_cancer	pancreas
parotid gland	salivary gland
pelvic LN	lymph node
penis	other genital
pericardium	mediastinum
peritoneum	peritoneum
pharynx	head and neck
pineal gland	brain
pituitary	brain
pleura	pleura
prostate	prostate
prostate_cancer	prostate
rectum	colorectum
retroperitoneal LN	lymph node
retroperitoneum	peritoneum
salivary	salivary gland
skeletal muscle	soft tissue
skin body	skin
skin lower face	skin
skin upper face	skin
skin	skin
small intestine	gastrointestinal tract
soft_tissue	soft tissue
spinal cord	central nervous system
spleen	other/unknown

sternoclavicular LN	lymph node
stomach	stomach
stomach_bowel	gastrointestinal tract
testes	testis
testicular_cancer	testis
testis	testis
thoracic LN	lymph node
thyroid	thyroid
tongue	oral cavity
tonsil	head and neck
trachea	upper respiratory tract
unkown primary	unknown primary
ureter	urinary tract
ureter_urethra	urinary tract
urinary_tract_cancer	urinary tract
uterus	uterus
vagina	other genital
vulva	other genital
adrenal	adrenal gland
lymphoma	lymphoma
leukemia	leukemia

Network-based analysis of cancer mortality misclassification reveals signatures of metastatic pathways: insights into breast-to-lung and other pairs of anatomical sites

Levy R. W. Cavalcante[a], Erneson A. Oliveira[a,b,c], Hermano A. L. Rocha[d], Antonio S. Lima Neto[c,e,f], Higor S. Monteiro[g], Rute M. Oliveira[g], Cesar I. N. Sampaio Filho[g], Luciano P. G. Cavalcanti[d,h], Tânia M. S. Coelho[f], Caterina A. M. La Porta[i,j,k], Stefano Zapperi[i,l,m], Humberto A. Carmona[g], José S. Andrade Jr.[g]

[a] *Postgraduate Program in Applied Informatics, University of Fortaleza, Fortaleza, 60811-905, Brazil*
[b] *Data Science and Artificial Intelligence Center, University of Fortaleza, Fortaleza, 60811-905, Brazil*
[c] *Professional Masters in City Science, University of Fortaleza, Fortaleza, 60811-905, Brazil*
[d] *Department of Community Health, Federal University of Ceará, Fortaleza, 60430-235, Brazil*
[e] *Center of Health Sciences, University of Fortaleza, Fortaleza, 60811-905, Brazil*
[f] *Health Secretariat of Ceará State, Fortaleza, 60060-440, Brazil*
[g] *Department of Physics, Federal University of Ceará, Fortaleza, 60455-760, Brazil*
[h] *Faculty of medicine, Christus University Center, Fortaleza, 60160-230, Brazil*
[i] *Center for Complexity and Biosystems, University of Milan, Milan, 20133, Italy*
[j] *Department of Environmental Science and Policy, University of Milan, Milan, 20133, Italy*
[k] *Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, 20122, Italy*
[l] *Department of Physics, University of Milan, Milan, 20133, Italy*
[m] *Consiglio Nazionale delle Ricerche, Istituto di Chimica della Materia Condensata e di Tecnologie per l'Energia, Milan, 20125, Italy*

Abstract

Background. Accurate cancer Cause-of-Death (CoD) data are fundamental for public health, but misclassification can bias mortality profiles, and existing correction methods fail to capture patient-level diagnostic trajectories. We propose a network-based framework to characterize CoD misclassification linked to metastatic spread.

Methods. We conducted a retrospective population-based study in Ceará, Brazil (2014–2019), linking mortality and cancer registry records. After filtering, 11,226 histologically confirmed cancer deaths were analyzed. We constructed a directed weighted Inconsistency Network (IN) between Last Diagnosis (LD) and CoD. Misclassification probabilities were compared with autopsy-derived metastasis probabilities (1,922 records), and effective mortality rates accounted

Email address: erneson@unifor.br (Erneson A. Oliveira)

for misclassification flows.
Findings. Overall, 22% of deaths showed no agreement between LD and CoD. Misclassification was not random: lung cancer was a major category into which deaths were incorrectly assigned in both sexes, whereas esophageal cancer showed this pattern only in males. The INs indicate reduced lung cancer mortality and increased stomach cancer mortality. In males, esophageal cancer mortality should also be reduced. In females, reduced lung cancer mortality is largely driven by breast-to-lung inconsistencies. This is compatible with autopsy evidence, suggesting that many deaths recorded as lung cancer should have been attributed to metastatic breast cancer.
Interpretation. Cancer mortality misclassification appears to follow plausible metastatic pathways. Pulmonary metastasis recorded as primary tumours may partly inflate lung cancer mortality while underestimating deaths from other primary tumours, notably breast cancer. This may bias surveillance, resource allocation, and cancer control policies. The framework offers a scalable approach to improve mortality estimates.
Funding. Brazilian agencies CNPq, CAPES, FUNCAP, and FEQ.

Keywords: Cause-of-death misclassification, Cancer mortality, Complex Networks, Oncology surveillance

---

Research in context

Evidence before this study

We searched PubMed/MEDLINE, SciELO, LILACS, Scopus, and Web of Science for studies published between Jan 1, 2000, and May 1, 2025. Search terms combined cancer sites and neoplasms with mortality, cause-of-death certification, vital statistics, garbage codes, misclassification, under-reporting, autopsy validation, verbal autopsy, minimally invasive tissue sampling, directed network models, graph theory, spatial analysis, time-series models, Brazil, Latin America, and low-income and middle-income countries. The search strategy used Boolean combinations including: ("colorectal cancer" OR "breast cancer" OR "lung cancer" OR "prostate cancer" OR "stomach cancer" OR "neoplasms") AND ("mortality" OR "cause of death" OR "death certificate" OR "vital statistics" OR "garbage codes") AND ("misclassification" OR "diagnostic accuracy" OR "metastatic attribution" OR "under-reporting") AND ("autopsy" OR "verbal autopsy" OR "post-mortem validation") AND ("directed network models" OR "graph theory" OR "spatial analysis"). Prior evidence indicates substantial bias in cancer mortality analyses based solely on uncorrected death certificates, especially in settings with regional heterogeneity in certification quality and persistent ill-defined causes. Existing WHO and GBD redistribution methods improve aggregate estimates but do not reconstruct individual diagnostic trajectories. No previous population-based study used directed network models to map biopsy-confirmed cancer diagnoses to administrative causes of death.

Added Value of This Study

This study adds a patient-trajectory perspective to cancer mortality surveillance by treating cause-of-death misclassification as a directed network rather than as random noise or an aggregate redistribution problem. Linking mortality records with population-based cancer registry data in Ceará, Brazil, we analysed 11,226 histologically confirmed cancer deaths from 2014 to 2019 and found that 22% had no recorded cancer diagnosis matching the administrative cause of death. The Inconsistency Network (IN) identified specific anatomial flows, showing that misclassification was directional and concentrated in clinically plausible pathways. In women, deaths with breast cancer as the last diagnosis were frequently certified as lung cancer, and the Breast–to–Lung pair showed a high inconsistency probability and an even higher metastasis probability in a curated autopsy dataset. In men, stomach cancer was frequently recorded as esophageal cancer, suggesting that upper gastrointestinal mortality rankings are sensitive to anatomical boundary and coding uncertainty. The Cervix-to-Uterus edge was very frequent but most certainly reflects contiguity or direct extension rather than distant metastatic attribution (metastasis probability close to zero). By translating these flows into effective mortality rates, the study shows how linked routine data and external validation evidence can reveal hidden distortions in cancer burden estimates and support more comparable surveillance across health systems with differing diagnostic capacity needs globally.

Implications of all the available evidence

All available evidence indicates that cancer mortality statistics should be interpreted as measurements that can contain systematic, site-specific error, not as neutral counts of underlying disease burden. This matters globally. GLOBOCAN 2022 estimated approximately 20 million new cancer cases and 9·7 million cancer deaths worldwide, and the cancers implicated in this study, lung, breast, stomach, and esophagus, are major contributors to global mortality or have pronounced geographic heterogeneity. The problem is not confined to settings with constrained diagnostic resources: evidence from Belgium has shown discordance between death certificates and medical files among patients with breast cancer, including missed breast-cancer deaths when metastatic sites or immediate causes were recorded instead. Health systems that rely on uncorrected death certificates may therefore mis-rank cancer priorities, overstate mortality assigned to terminal metastatic sites, and underestimate the primary tumours driving fatal trajectories. The findings support routine linkage of cancer registries, mortality systems, and hospital records; targeted review of high-risk anatomical pairs such as Breast-to-Lung and Stomach-to-Esophagus; and sentinel validation through autopsy, minimally invasive tissue sampling, or expert clinical review. Used across countries, the IN could strengthen international comparability and make cancer-control planning more equitable, data-responsive, transparent, auditable, and clinically interpretable for decision-makers and patients worldwide alike.

Introduction

Cancer surveillance underpins public-health policy because accurate, cause-specific mortality data guide burden estimation, priority setting, resource allocation, and evaluation of prevention and treatment programmes. Administrative mortality records are useful, however, only when Cause-of-Death (CoD) assignments are valid.[1,2] In Brazil, the mortality information system offers universal coverage, but regional variation in certification practices and persistent ill-defined causes weaken mortality fractions used for planning.[1] In administrative and electronic health record datasets, misclassification can bias estimates unless corrected with validation data.[2] Surveillance methods must therefore account for death-certificate error and incorporate independent reference information to support equitable cancer control policy.[2,3]

CoD misclassification is common in routine certificates and Verbal Autopsy (VA), varying by cause and context. Physician-certified VAs and automated VA algorithms show heterogeneous sensitivity and chance-corrected agreement, performing better for chronic conditions than for many infectious and non-specific causes.[1,3] Autopsy comparisons show concordance for selected conditions but low predictive values for others, confirming systematic and context-dependent errors that distort Cause-Specific Mortality Fractions (CSMFs) and may produce spurious trends without correction.[1,2,4]

Post-mortem examination therefore remains essential. Complete diagnostic autopsy is the most definitive standard for identifying the pathological processes responsible for death and validating CoD assignments from clinical records or VAs.[3,5] However, declining autopsy rates, biosafety concerns, cultural constraints, and resource limitations have encouraged alternatives, including Minimally Invasive Tissue Sampling (MITS), targeted post-mortem sampling, and Post-Mortem Imaging (PMI) with tissue sampling. Evaluations indicate that these methods can increase agreement with autopsy and provide scalable validation datasets where full autopsy is impractical.[6–8] Nevertheless, PMI alone may miss vascular perfusion disorders and subtle histopathological findings. Robust CoD ascertainment often requires combining autopsy, image-guided sampling, and clinical data.[5,6,9]

This issue is relevant to cancer surveillance in Brazil and other settings, where coding, diagnostic intensity, and certification practices may change faster than disease dynamics. Such changes can create apparent paradoxes, such as rising recorded lung-cancer mortality despite improved diagnostic and therapeutic capacity. Without independent post-mortem validation, it is difficult to distinguish true increases in fatal disease from artifacts of ascertainment, coding, or misclassification.[1,2] VA and other routine methods perform unevenly across cancer types and settings, while autopsy- or MITS-based verification may reveal under- and over-attribution of specific causes.[1,4,7] Thus, an "improved care with increased mortality" signal may reflect a genuine epidemiological shift, a certification-driven artifact, or both. Resolving this ambiguity requires linkage between administrative surveillance and post-mortem evidence.[1,2,6]

To address this gap, we designed a retrospective, population-based study of

cancer mortality in Ceará, Brazil, integrating administrative health data and population estimates for 2014–2019 with a curated autopsy dataset from the literature and validated post-mortem series. This design uses validation samples to correct misclassified outcomes and re-estimate CSMFs with reduced bias,[2] while building on Brazilian evidence that autopsy-anchored validation improves the interpretability of mortality statistics.[1] We include image-guided and MITS-derived sources where full autopsy is unavailable, based on PMCT and targeted post-mortem sampling studies.[6–9] Our objective is to quantify the extent and direction of misclassification in recorded cancer deaths, clarify trends such as lung cancer in Ceará, and produce corrected mortality estimates for cancer-control planning.[1,2,6]

Methods

Study design

We conducted a retrospective population-based observational study of cancer in the state of Ceará, Brazil, using linked administrative health data during the study period from 2014 to 2019. Additional mortality records and population estimates at both the state and national levels were also considered to extend the analyses from 2009 to 2023. We also incorporated a manually curated autopsy dataset of cancer-related deaths compiled from the literature. All study procedures were approved by the Research Ethics Committee of Centro Universitário Christus - Unichristus (CEP/Unichristus; approval no. 5.617.938; CAAE 30203820.8.0000.5049)

Study Setting

Ceará, a state in Northeast Brazil, had a population of approximately 9·2 million in the 2022 census conducted by the Brazilian Institute of Geography and Statistics (Instituto Brasileiro de Geografia e Estatística [IBGE]).[10] Compared with Brazil overall, Ceará has a younger population structure and lower average income and educational attainment, reflecting persistent regional socioeconomic inequalities. Individuals aged 60 years or older account for approximately 15–16% of the state population, consistent with the ongoing demographic ageing observed nationally. Life expectancy at birth in Ceará (76·9 years) is slightly higher than the national average (75·5 years in 2022), although heterogeneity persists within the state. Mortality patterns are consistent with the national epidemiological transition, with non-communicable diseases accounting for most deaths, alongside a continued contribution from external causes and communicable diseases.[10]

Datasets

We used four datasets in this study: the Registry Dataset, Mortality Dataset, Population Dataset, and Autopsy Dataset. The Registry Dataset comprised a final sample of 11,226 individuals from Ceará, Brazil, obtained through linkage between the Mortality Information System (Sistema de Informações sobre

Mortalidade [SIM])[11] and the Population-Based Cancer Registries (Registros de Câncer de Base Populacional [RCBP])[12], including information on CoD and longitudinal cancer diagnostic histories for individuals who died between 2014 and 2019. The Mortality Dataset comprised 128,863 cancer death records from Ceará and 3,200,318 from Brazil spanning 2009–2023, extracted from the SIM. The Population Dataset included annual population estimates for all Brazilian states and the Federal District from 2009 to 2023, obtained from the IBGE[13]. Finally, the Autopsy Dataset, compiled by Font-Clos and colleagues,[14] comprised 1,922 anonymised records harmonised from five independent autopsy studies published between 1950 and 2016, containing information on primary cancer sites and metastatic dissemination patterns. Detailed descriptions of the datasets, inclusion and exclusion criteria, variable definitions, and data harmonisation procedures are provided in the Datasets section of the Supplementary Materials (SM).

Consistent and inconsistent deaths

In the Registry Dataset, we classify deaths according to a conservative definition of agreement between the Diagnostic History and the CoD. A consistent death is defined as a death in which the patient has at least one diagnosis recorded in the Diagnostic History that matches the CoD, whereas an inconsistent death is defined as a death in which the patient has no diagnosis recorded in the Diagnostic History that matches the CoD. In this context, we define N^{con} as the total number of consistent deaths, N^{inc} as the total number of inconsistent deaths, and $N = N^{\text{con}} + N^{\text{inc}}$ as the total number of deaths.

The Model

We propose an Inconsistency Network (IN) model to characterize discrepancies between the Diagnostic History and the CoD recorded in the Registry Dataset. The IN is defined as a directed weighted graph in which each node i represents an anatomical site associated with a malignant neoplasm, identified through broad organ classes. A directed edge from node i to node j is established when the anatomical site associated with the patient's Last Diagnosis (LD) of a malignant primary tumor differs from the anatomical site recorded as the CoD, with i denoting the LD site and j the CoD site. Thus, edges represent only inconsistent site assignments between these two records. The weight w_{ij} of the directed edge (i, j) is defined as the total number of inconsistent deaths from i to j, that is, the number of patients whose LD was assigned to site i while the CoD was assigned to a different site j. In this way, the IN summarizes how frequently deaths attributed to one cancer site in mortality records are associated, in the linked diagnostic history, with a different site of malignant neoplasm. We assume that the LD provides the best available proxy for the patient's disease when no earlier diagnosis in the diagnostic history matches the CoD. The biological plausibility of this assumption is examined later in the discussion.

Mortality rate

The mortality rate of node i in year t is computed as,

$$\mu_i(t) = \frac{n_i(t)}{p(t)} \times 100{,}000 \text{ inhabitants}, \tag{1}$$

where $n_i(t)$ is the number of deaths of the node i in year t and $p(t)$ is the population estimate in the same year t. For Ceará and Brazil, the values of $n_i(t)$ and $p(t)$ are obtained from the Mortality and Population Datasets, respectively.

Effective mortality rate

We define the effective mortality rate of node i in year t as,

$$\mu_i^{\text{eff}}(t) = \alpha_i \mu_i(t), \tag{2}$$

with α_i representing an adjustment factor that accounts for the redistribution of misclassified deaths. This factor is given by,

$$\alpha_i = \frac{n_i^{\text{eff}}}{n_i}, \tag{3}$$

where n_i^{eff} is the effective number of deaths of the node i, defined as,

$$n_i^{\text{eff}} = n_i - s_i^{\text{in}} + s_i^{\text{out}}, \tag{4}$$

with $s_i^{\text{in}} = \sum_j w_{ji}$ and $s_i^{\text{out}} = \sum_j w_{ij}$ representing the in-strength (or weighted in-degree) and out-strength (or weighted out-degree) of the node i, respectively.

Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

Table 1 summarizes the Registry dataset with respect to the diagnostic history lengths ℓ. We found that 2,502 ($\sim 22\%$) deaths are inconsistent, i.e., they have no diagnoses that match the recorded CoD. These inconsistencies occur across all values of ℓ, exhibiting an almost equal distribution between males and females. Among these inconsistent deaths, only 33 different CoD categories were observed (see Tab. 2). The most frequent CoD include lung cancer, gastrointestinal tract, stomach cancer, and esophageal cancer. Regarding the LD, Tab. 3 shows the 28 categories observed among inconsistent deaths. The distribution is dominated by skin cancers, stomach cancer, colorectum cancer, and cervix cancer. Breast cancer is observed almost exclusively among females, with only two male records in CoD and eight in LD. Finally, Tab. 4 reorganizes these results by focusing on the distribution of LD–to-CoD pairs among inconsistent deaths, stratified by sex, rather than considering LD and CoD separately. These

patterns suggest that the inconsistencies tend to emerge in specific directed LD-to-CoD pairs, indicating a non-random misclassification across CoD categories.

We show the combined, male and female INs in Fig. 1. A core subnetwork was highlighted composed of the anatomical sites that cumulatively account for 50% of all inconsistent deaths, N^{inc}. In Fig. 1A, Lung, Breast, and Stomach account for the largest numbers of deaths for both sexes. We identify three prominent edges: Cervix-to-Uterus, Breast-to-Lung, and Stomach-to-Esophagus. In Fig. 1B, the largest nodes correspond to Stomach, Lung, and Prostate, indicating that these broad organ classes account for the highest numbers of deaths. The adjustment factors α_i reveal that mortality rates μ_i for esophageal cancer should be reduced ($\alpha_i < 1$), whereas μ_i for stomach cancer should be increased ($\alpha_i > 1$). The most heavily weighted (thickest) edge in the male network connects stomach cancer to esophageal cancer, indicating that indeed a substantial number of deaths with stomach cancer in LD were officially recorded as esophageal cancer in CoD. In Fig. 1C, the largest nodes are breast, lung, and cervix cancers. In this case, lung cancer presents $\alpha_i < 1$, indicating that its mortality rate μ_i should be reduced, while breast cancer presents $\alpha_i > 1$, indicating that its μ_i should be increased. Excluding the Cervix-to-Uterus edge, the most heavily weighted edge is Breast-to-Lung, indicating that inconsistencies involving breast cancer are the main driver of the reduction in the μ_i^{eff} for lung cancer among females.

From the Mortality Dataset described in Methods, Fig. 2 shows the time series of mortality rates μ_i for selected cancers in Brazil and Ceará, along with the corresponding effective mortality rates μ_i^{eff} for Ceará. In Figs. 2A and 2B, we show that μ_i^{eff} for stomach cancer is consistently higher than the corresponding μ_i in Ceará, whereas for esophageal cancer it is systematically lower. This adjustment amplifies the contrast between the state and national patterns, reinforcing the divergence between Ceará and Brazil for these male cancers over the studied period. For females (Figs. 2C and 2D), μ_i^{eff} for breast cancer is higher than the reported μ_i in Ceará, while for lung cancer it is lower. In this case, the adjustment leads to a closer alignment between Ceará and Brazil, particularly by reducing the gap in lung cancer mortality and increasing the comparability of breast cancer trends with national levels.

Figure 3 relates the inconsistency probability P_{ij}^{inc} (the probability that a death from cancer site i is misclassified as site j) to the metastasis probability P_{ij}^{met} (the probability of metastasis from primary cancer site i to metastatic site j) on a logarithmic scale (see details about their calculation in the SM). Each point represents a cancer site pair, with filled blue circles highlighting those belonging to the core subnetwork and empty circles representing all other pairs. Among the notable pairs from the core subnetwork, the Breast-to-Lung pair stands out by exhibiting a high value for P_{ij}^{inc} and an even higher value for P_{ij}^{met}, suggesting that such inconsistencies may be indeed associated with unreported metastatic deaths.

Finally, we show the Wilson score confidence intervals CI_{WS} (see SM) for the previous highlighted pairs in Fig. 4. For each pair, P_{ij}^{inc} (green markers) and P_{ij}^{met} (orange markers) are displayed together with their corresponding two-

sided horizontal error bars. These intervals allow a direct assessment of both magnitude and statistical uncertainty. Consistent with the pattern observed in Fig. 3, the Breast-to-Lung pair exhibits high values for both probabilities along with narrow values of CI_{WS}, indicating high precision in our estimation. Together, these values suggest that a substantial fraction of deaths recorded as lung cancer in CoD but preceded by breast cancer in LD may correspond to metastatic progression rather than true misclassification.

Discussion

This retrospective, population-based observational study of cancer-related mortality in Ceará, Brazil, found that about 22% of 11,226 histologically confirmed cancer deaths were inconsistent: the recorded CoD did not match any cancer diagnosis in the patient's history. The IN analysis showed directional over-attribution to lung cancer, particularly among women, and to esophageal cancer among men. After redistribution, effective mortality rates decreased for female lung cancer and male esophageal cancer, but increased for female breast cancer and male stomach cancer. Stomach cancer had adjustment factors above unity, indicating systematic under-recording. The Breast-to-Lung pair showed high inconsistency and metastasis probabilities with narrow Wilson confidence intervals, suggesting that many deaths recorded as lung cancer in women with prior breast cancer may represent unreported metastatic progression rather than primary lung malignancy. Thus, cancer CoD misclassification appears not merely random, but structured and directional. Where mortality rankings guide cancer-control priorities, such errors may overstate deaths assigned to terminal metastatic sites, particularly lung, while understating the primary tumours that initiated fatal trajectories, notably breast cancer in women and stomach cancer in men.

These findings align with evidence that routine CoD certification is systematically error-prone. In São Paulo, a validation study of 2,060 decedents found that physician-certified VA varied by cause when compared with conventional autopsy: population-level CSMF accuracy was 0·81, but chance-corrected accuracy was only 0·49.[1] Our study extends this evidence to Northeast Brazil, where socioeconomic vulnerability and lower diagnostic intensity may amplify misclassification embedded in clinical histories and death certificates. A systematic review found only two validation studies using full autopsy as the gold standard, with opposite conclusions.[3] In sub-Saharan Africa, Interpreting VA (InterVA) showed only 22% positive predictive value against complete autopsy, and 5·3% for communicable diseases, confirming poor individual-level reliability even when population-level CSMFs appear acceptable.[4] The IN approach complements these studies by making misclassification direction explicit and linking it to biologically plausible metastatic pathways.

The over-recording of lung cancer in Ceará, especially the Breast-to-Lung edge among women, matters globally. Global Burden of Disease data showed that lung cancer deaths increased by 91·8% from 1990–2019, largely because of

ageing and population growth, while age-standardised mortality declined non-significantly. Mortality sources include vital registration, verbal autopsy, and registries, each with ascertainment bias.[5] Rising recorded lung cancer mortality may therefore combine true epidemiological change with ascertainment artefact. Other high-weight edges clarify the network's clinical meaning. The Cervix-to-Uterus pair, the most frequent inconsistent LD-to-CoD pair, likely reflects anatomical contiguity, coding granularity, or direct extension rather than distant metastasis.[15,16] By contrast, the Stomach-to-Esophagus edge, especially among men, reflects classification challenges across the anatomically continuous stomach, cardia, distal esophagus, and esophagogastric junction. The decrease in effective esophageal cancer mortality and increase in effective stomach cancer mortality suggest that upper gastrointestinal rankings are sensitive to site-attribution errors, a global concern because stomach cancer remains a leading cancer killer and esophageal cancer has strong geographic heterogeneity.[17]

The IN framework helps disentangle these effects. After redistributing deaths misclassified from breast to lung cancer, the female gap between Ceará and national lung cancer trends narrows, whereas the breast cancer gap widens, suggesting that part of the apparent lung burden reflects unrecognised metastatic breast cancer. This is biologically coherent because breast cancer commonly metastasises to lung, and the autopsy dataset from five studies showed high Breast-to-Lung metastasis probability. The estimators proposed by Shen and colleagues,[2] which combine large misclassified databases with smaller validation samples subject to selection bias, support this hybrid design. Brazilian surveillance studies reinforce linked data: Galvão and colleagues[18] described challenges integrating population-based registries, hospital registries, and SIM in Mato Grosso, while Soares and colleagues[19] showed that leukaemia and lymphoma mortality trends differed depending on state versus national data.

The implications extend beyond Ceará because international mortality comparisons depend on accurate primary-site attribution. GLOBOCAN 2022 estimated about 20 million new cancer cases and 9·7 million cancer deaths worldwide, with lung, colorectal, liver, breast, and stomach cancers among the leading causes of cancer mortality.[20] Even modest directional misclassification between primary and terminal sites can alter rankings. This is not confined to LMICs: in Belgium, discordance between death certificates and medical files among breast cancer patients showed that breast cancer deaths may be missed when metastatic sites, or immediate causes, are recorded instead of the underlying primary malignancy.[21] The Breast-to-Lung pattern is globally relevant: lung cancer is the leading cause of cancer mortality worldwide, while breast cancer may be underestimated when pulmonary metastases are certified as primary lung cancer. The IN framework should therefore be viewed as a transferable surveillance method for detecting site-specific distortions in cancer mortality data.

Implications for Public Health and Policy

Policy implications are substantial and direct. If metastatic breast cancer deaths are assigned to lung cancer, unadjusted SIM data may overestimate pri-

mary lung cancer and misallocate resources toward lung cancer screening, early detection, and treatment, at the expense of breast cancer control. These results do not reduce the importance of tobacco control or lung cancer services. Rather, they show that uncorrected rankings may miscalibrate primary-site priorities. Access to diagnosis and treatment strongly affects survival: in Goiânia, five-year survival for metastatic breast cancer was 6·9% in the public system versus 22·4% in the private sector.[22] If Ceará's true breast cancer burden is higher than recorded, expanding public diagnosis and treatment is more urgent. Likewise, under-recorded stomach cancer may lead to insufficient investment in Helicobacter pylori eradication, endoscopic surveillance, and gastric surgery. Regional SIM quality is uneven, with the Northeast historically showing higher ill-defined CoD proportions,[1] and this study shows that such unevenness produces directional cancer-type misclassification.

Misclassification also affects evaluation of cancer-control programmes. Inflated lung cancer mortality may distort tobacco-control assessments, whereas under-recorded stomach cancer may mask gastric cancer trends. Similar correction needs exist elsewhere. The American Cancer Society adjusted mortality rates for American Indian/Alaska Native populations because racial misclassification on death certificates underestimated the true burden.[23] Linked registry–mortality analyses are therefore essential for planning,[18] and the IN framework offers a concrete method for implementing them. A Belém cervical-cancer registry study found decreasing age-adjusted incidence from 1998–2017 but stable mortality, a pattern that could reflect true stability or misclassification. The IN framework could help resolve such ambiguities.[24]

Strengths and Limitations

Strengths include the large population-representative dataset, linkage to SIM and RCBP, histologically confirmed diagnoses, transparent IN modelling, autopsy-derived metastasis probabilities, and Wilson confidence intervals for uncertainty. Limitations include non-contemporary metastasis estimates, absence of information on stage, treatment, and diagnosis-to-death interval as well as selection bias from histological confirmation, limited generalisability beyond Ceará, and heterogeneity in the autopsy data.

Recommendations for Action

Recommended actions include systematic state-level linkage of SIM, RCBP, and hospital cancer registries using IN analysis to monitor directional misclassification, consistent with calls to reduce incompleteness and improve cause-specific mortality data.[18,23,25] Brazilian National Cancer Institute (Instituto Nacional de Câncer [INCA]) should consider sentinel autopsy or MITS in representative cancer deaths; PMCT plus targeted sampling improves agreement with autopsy-determined CoD, and MITS is scalable and cost-effective.[6–8] Death-certificate training should emphasize the distinction between primary site, metastatic site, direct extension, and terminal events, including Breast-to-Lung, Stomach-to-Esophagus, and Cervix-to-Uterus scenarios. Finally, cancer-control evaluations

should incorporate linked registry–mortality analyses and, where feasible, sentinel autopsy or MITS to generate corrected mortality estimates that are comparable across regions and over time.

Contributors

EAO, HALR, ASL, CAMLP, SZ, HAC, and JSA conceptualized the study. EAO, HALR, CAMLP, SZ, HAC, and JSA developed the methodology. LRWC, EAO, HALR, HSM, CAMLP, SZ, HAC, and JSA contributed to the formal analysis and data curation. LRWC and EAO were responsible for the software and visualization. EAO, HALR, ASL, HSM, RMO, CINS, LPGC, TMSC, CAMLP, SZ, HAC, and JSA contributed to the interpretation and discussion of the results. LRWC, EAO, HALR, and JSA wrote the manuscript. JSA was responsible for supervision and funding acquisition. All authors reviewed the manuscript and approved the final version for submission.

Declaration of Interests

The authors declare no competing interests.

Data sharing statement

All datasets and codes used in this study are available at `https://github.com/erneson/inconsistency_networks`, except for the Autopsy Dataset, which is available at `https://github.com/ComplexityBiosystems/CTC-model`.

Acknowledgements

We gratefully acknowledge CNPq, CAPES, FUNCAP, the National Institute of Science and Technology for Complex Systems in Brazil and the Edson Queiroz Foundation for financial support. SZ and CAMLP also acknowledge specific funding support: SZ from CAPES (PrInt, process n. 88887.937759/2024-00), and CAMLP from CAPES (PrInt, process n. 88887.937852/2024-00) and AIRC, Next Generation Clinician Scientist 2022 (ID 28402, project, PI Carlotta Palumbo). JSA acknowledges funding from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Award 303765/2017-8, Bolsa PQ).

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the authors used ChatGPT in order to perform language editing, including the correction of grammatical errors, spelling mistakes, and minor improvements in writing clarity. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

[1] de André CDS, Bierrenbach AL, Barroso LP, et al. Validation of physician certified verbal autopsy using conventional autopsy: a large study of adult non-external causes of death in a metropolitan area in Brazil. BMC Public Health. 2022;22(1):748.

[2] Shen J, Isenberg D, Linn K, Hubbard R. Integrating Misclassified EHR Outcomes With Validated Outcomes From a Non-Probability Sample. Stat Med. 2025;44(15-17):e70127.

[3] Bailo P, Gibelli F, Ricci G, Sirignano A. Verbal Autopsy as a Tool for Defining Causes of Death in Specific Healthcare Contexts: Study of Applicability through a Traditional Literature Review. Int J Environ Res Public Health. 2022;19(18):11749.

[4] Yokobori Y, Matsuura J, Sugiura Y, et al. Comparison of the Causes of Death Identified Using Automated Verbal Autopsy and Complete Autopsy among Brought-in-Dead Cases at a Tertiary Hospital in Sub-Sahara Africa. Appl Clin Inform. 2022;13(03):583-91.

[5] Michaud K, Jacobsen C, Basso C, et al. Application of postmortem imaging modalities in cases of sudden death due to cardiovascular diseases–current achievements and limitations from a pathology perspective. Virchows Arch. 2022;482(2):385-406.

[6] Mentink MG, Latten BG, Bakers FC, et al. Efficacy of postmortem CT and tissue sampling in establishing the cause of death in clinical practice: a prospective observational study. J Clin Pathol. 2022;77(4):259-65.

[7] Toit J, Storath K, Dunn I, et al. Experience on postmortem minimally invasive tissue sampling to ascertain the cause of death determination in South African children: A case for implementing as standard of care. S Afr Med J. 2024;114(2):e1538.

[8] Peñuelas N. Diagnostic accuracy of a simplified minimally invasive tissue sampling protocol for stillbirths in low-resource settings. BMJ Glob Health. 2025;10(10):e018380.

[9] Zirn A, Scheurer E, Lenz C. Automated detection of fatal cerebral haemorrhage in postmortem CT data. Int J Legal Med. 2024;138(4):1391-9.

[10] Instituto Brasileiro de Geografia e Estatística (IBGE). Censo Demográfico 2022; 2026. `https://censo2022.ibge.gov.br/` (accessed Jun 15, 2026).

[11] Ministério da Saúde (MS) do Brasil. Sistema de Informação sobre Mortalidade (SIM); 2026. `http://sim.saude.gov.br/default.asp` (accessed Jun 15, 2026).

[12] Instituto Nacional de Câncer (INCA). Registro de Câncer de Base Populacional (RCBP); 2026. https://www.inca.gov.br/en/numeros-de-cancer/registro-de-cancer-de-base-populacional (accessed Jun 15, 2026).

[13] Instituto Brasileiro de Geografia e Estatística (IBGE). Projeções da População: Brasil e Unidades da Federação; 2026. https://www.ibge.gov.br/estatisticas/sociais/populacao/9109-projecao-da-populacao.html (accessed Jun 15, 2026).

[14] Font-Clos F, Zapperi S, La Porta CA. Blood flow contributions to cancer metastasis. Iscience. 2020;23(5).

[15] He F, Li W, Liu P, et al. Influence of uterine corpus invasion on prognosis in stage IA2–IIB cervical cancer: A multicenter retrospective cohort study. Gynecol Oncol. 2020;158(2):273-81.

[16] Bhatla N, Berek JS, Cuello Fredes M, et al. Revised FIGO staging for carcinoma of the cervix uteri. Int J Gynaecol Obstet. 2019;145(1):129-35.

[17] Oh SE, Park S, Ahn S, et al. Prognostic significance of esophagogastric junction invasion in patients with adenocarcinoma of the cardia or subcardia. Cancers. 2023;15(6):1656.

[18] Galvão ND, Souza RAGd, Souza BdSNd, et al. Cancer surveillance in Mato Grosso, Brazil: methodological and operational aspects of a university extension/research project. Rev Bras Epidemiol. 2022;25(Suppl 1):e220002.

[19] Soares MR, Melanda FN, Lima Neto GSd, et al. Mortality trend and analysis of potential years of life lost due to leukemia and lymphoma in Brazil and Mato Grosso. Rev Bras Epidemiol. 2022;25(Suppl 1):e220008.

[20] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. Ca Cancer J Clin. 2018;68(6):394-424.

[21] Izci H, Tambuyzer T, Vandeven J, et al. Cause of death for patients with breast cancer: discordance between death certificates and medical files, and impact on survival estimates. Arch Public Health. 2021;79(1):111.

[22] Soares LR, Freitas-Junior R, Curado MP, Paulinelli RR, Martins E, Oliveira JC. Low overall survival in women with de novo metastatic breast cancer: does this reflect tumor biology or a lack of access to health care? JCO Global Oncology. 2020;6:679-87.

[23] Islami F, Baeker Bispo J, Lee H, et al. American Cancer Society's report on the status of cancer disparities in the United States, 2023. CA Cancer J Clin. 2024;74(2):136-66.

[24] Formigosa LAC, Dos Santos LF, Formigosa JDC, da Silva Pinto D, Dos Reis DST, da Silva MVS. Impact of screening on cervical cancer incidence and mortality in a Northern Brazilian city. ecancermedicalscience. 2022;16:1418.

[25] Fan Y, Jiang Y, Gong L, et al. Epidemiological and demographic drivers of lung cancer mortality from 1990 to 2019: results from the global burden of disease study 2019. Front Public Health. 2023;11:1054200.

ℓ	N	N^{con}	N^{inc}	$N^{\text{con}}_{\text{male}}$	$N^{\text{con}}_{\text{female}}$	$N^{\text{inc}}_{\text{male}}$	$N^{\text{inc}}_{\text{female}}$
1	8936	6827	2109	3384	3443	1035	1074
2	1657	1367	290	713	654	160	130
3	450	377	73	204	173	44	29
4	130	109	21	49	60	16	5
5	36	30	6	18	12	4	2
6	11	9	2	5	4	2	0
7	4	4	0	3	1	0	0
8	1	0	1	0	0	1	0
9	1	1	0	0	1	0	0
Total	11226	8724	2502	4376	4348	1262	1240

Table 1: Overview of diagnostic history lengths ℓ in the Registry Dataset. For each ℓ, the table shows the total number of deaths N, the total number of consistent deaths N^{con} — when at least one diagnosis matches the Cause of Death (CoD) – and the total number of inconsistent deaths N^{inc} — when none do. N^{con} and N^{inc} are further stratified by sex: $N^{\text{con}}_{\text{male}}$, $N^{\text{con}}_{\text{female}}$, $N^{\text{inc}}_{\text{male}}$, and $N^{\text{inc}}_{\text{female}}$. Overall, we found that approximately 22% of the deaths have no diagnoses that match the recorded CoD, with an almost equal distribution between males and females.

Cause of Death (CoD)	N^{inc}	$N^{\text{inc}}_{\text{male}}$	$N^{\text{inc}}_{\text{female}}$
Lung	344	167	177
Gast. Tract	188	93	95
Stomach	180	108	72
Esophagus	165	129	36
Uterus	158	0	158
Liver	152	71	81
URT	122	106	16
Pancreas	102	50	52
Head & Neck	101	75	26
Lymphoma	96	48	48
Colorectum	89	42	47
Cervix	78	0	78
Prostate	78	78	0
Peritoneum	74	33	41
Bone	68	36	32
Brain	67	25	42
Biliary Tract	66	28	38
Breast	62	2	60
Ovary	52	0	52
Soft Tissue	40	25	15
CNS	36	17	19
Skin	33	21	12
Oral	25	20	5
Leukemia	25	17	8
Kidney	24	20	4
Bladder	23	16	7
Mediastinum	18	8	10
Other/Unknown	10	8	2
Other Genital	7	1	6
Salivary Gland	7	6	1
Urinary Tract	6	6	0
Testis	5	5	0
Pleura	1	1	0
Total	2502	1262	1240

Table 2: Cause of Death (CoD) in the Registry Dataset. The table shows the total number of inconsistent deaths N^{inc}, the total number of inconsistent deaths for male $N^{\text{inc}}_{\text{male}}$, and the total number of inconsistent deaths for female $N^{\text{inc}}_{\text{female}}$ for all CoD, aggregated into broad organ classes and listed in descending order of N^{inc}. Abbreviations used include Gast. Tract (Gastrointestinal Tract), URT (Upper Respiratory Tract), and CNS (Central Nervous System).

Last Diagnosis (LD)	N^{inc}	$N^{\text{inc}}_{\text{male}}$	$N^{\text{inc}}_{\text{female}}$
Skin	351	205	146
Stomach	287	191	96
Colorectum	234	113	121
Cervix	189	0	189
Oral	161	132	29
Breast	147	8	139
Lung	106	52	54
Prostate	92	92	0
Head & Neck	85	71	14
Esophagus	73	48	25
Liver	71	36	35
Mediastinum	71	31	40
Bladder	64	56	8
Ovary	62	0	62
Soft Tissue	60	24	36
URT	59	52	7
Biliary Tract	59	26	33
Gast. Tract	57	26	31
Uterus	52	0	52
Peritoneum	48	18	30
Brain	45	23	22
Other Genital	36	7	29
Pancreas	27	14	13
Bone	22	11	11
Kidney	22	12	10
Salivary Gland	18	11	7
CNS	2	1	1
Testis	2	2	0
Total	2502	1262	1240

Table 3: Last Diagnosis (LD) in the Registry Dataset. The table shows the total number of inconsistent deaths N^{inc}, the total number of inconsistent deaths for male $N^{\text{inc}}_{\text{male}}$, and the total number of inconsistent deaths for female $N^{\text{inc}}_{\text{female}}$ for all LDs, aggregated into broad organ classes and listed in descending order of N^{inc}. Abbreviations used include Gast. Tract (Gastrointestinal Tract), URT (Upper Respiratory Tract), and CNS (Central Nervous System).

Last Diagnosis (LD)	Cause of Death (CoD)	N^{inc}	$N^{\text{inc}}_{\text{male}}$	$N^{\text{inc}}_{\text{female}}$
Cervix	Uterus	115	0	115
Stomach	Esophagus	99	72	27
Colorectum	Gast. Tract	87	37	50
Skin	Lung	66	38	28
Stomach	Gast. Tract	58	34	24
Oral	Head & Neck	56	44	12
Breast	Lung	52	2	50
Oral	URT	46	42	4
Head & Neck	URT	46	38	8
Colorectum	Stomach	42	21	21
Esophagus	Stomach	40	25	15
Mediastinum	Lung	39	21	18
Stomach	Liver	30	24	6
Brain	CNS	26	12	14
Liver	Biliary Tract	26	11	15
Skin	Lymphoma	26	18	8
Oral	Lung	24	18	6
Bladder	Prostate	24	24	0
Biliary Tract	Liver	24	11	13
Skin	Stomach	23	16	7
URT	Head & Neck	23	19	4
Skin	Prostate	22	22	0
Colorectum	Lung	21	9	12
Colorectum	Liver	21	10	11
Ovary	Cervix	20	0	20
Total		1056	568	488

Table 4: Top 25 Last Diagnosis-to-Cause of Death (LD-to-CoD) pairs in the inconsistent deaths. The table shows the total number of inconsistent deaths N^{inc}, the total number of inconsistent deaths for male $N^{\text{inc}}_{\text{male}}$, and the total number of inconsistent deaths for female $N^{\text{inc}}_{\text{female}}$ for the most frequent LD-to-CoD pairs, aggregated into broad organ classes and listed in descending order of N^{inc}. Abbreviations used include Gast. Tract (Gastrointestinal Tract), URT (Upper Respiratory Tract), and CNS (Central Nervous System).

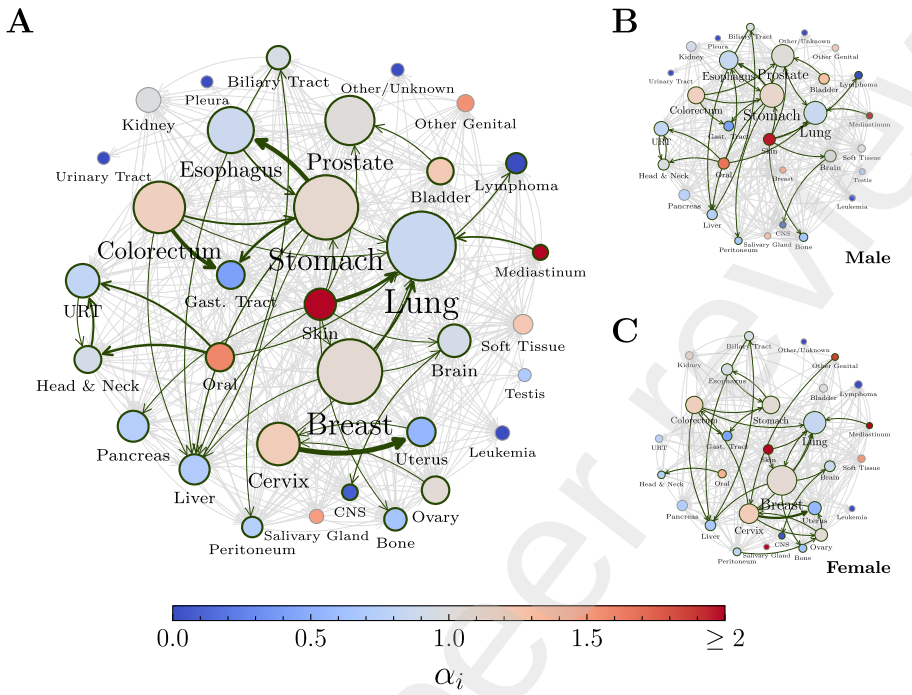

Figure 1: Combined, Male and Female Inconsistency Networks (INs). The networks summarize discordances between the anatomical site associated with the Last Diagnosis (LD) of a malignant primary tumor and the anatomical site recorded as the Cause of Death (CoD) in the Registry Dataset. Each node represents an anatomical site associated with a malignant neoplasm, identified through broad organ classes. Node size is proportional to the number of deaths assigned to the corresponding broad organ class as CoD. Node color represents the adjustment factor α_i, which quantifies the correction required to adjust mortality rates based on CoD using information from the IN. A directed edge from node i to node j indicates that the LD site was i whereas the CoD site was j, with $i \neq j$; thus, edges represent only inconsistent site assignments. Edge width is proportional to the edge weight w_{ij}, defined as the number of deaths for which the LD corresponds to site i and the CoD to site j. Nodes with a green border and their connecting green edges highlight a core subnetwork, representing the anatomical sites that cumulatively account for 50% of the total number of inconsistent deaths N^{inc}. (A) Combined IN for both sexes, in which Lung, Breast, and Stomach account for the largest numbers of deaths. We highlight three edges: Cervix-to-Uterus, Breast-to-Lung, and Stomach-to-Esophagus. While Cervix-to-Uterus is the most heavily weighted edge and likely reflects direct tumor invasion between adjacent organs rather than true misclassification, the Stomach-to-Esophagus edge may reflect classification challenges within the anatomically continuous esophagogastric region. In contrast, Breast–to–Lung emerges as a prominent and clinically plausible pattern of inconsistency. (B) Male IN, in which Stomach, Lung, and Prostate account for the largest numbers of deaths. The edge with the highest number of inconsistent deaths is Stomach-to-Esophagus. The corresponding α_i values indicate that mortality rates for Esophagus should be reduced, whereas those for Stomach should be increased. The reduction associated with Esophagus appears to be primarily driven by inconsistencies involving Stomach. (C) Female IN, in which Breast, Lung, and Cervix account for the largest numbers of deaths. In this case, excluding Cervix-to-Uterus, the highest-weight edge is Breast-to-Lung. The α_i values indicate that the mortality rate for Lung should be reduced, whereas that for Breast should be increased. Notably, the reduction in Lung is largely driven by inconsistencies involving Breast. Abbreviations used include Gast. Tract (Gastrointestinal Tract), URT (Upper Respiratory Tract), and CNS (Central Nervous System)

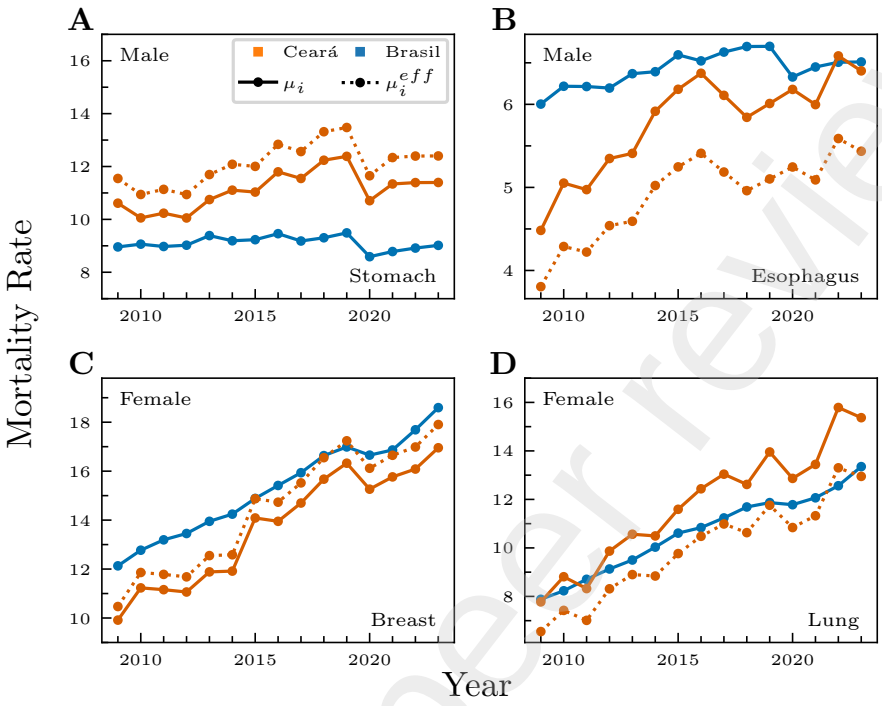

Figure 2: Time series of mortality rates (per 100,000 inhabitants) for selected cancers in Brazil and Ceará from 2009 to 2023. (A) Male mortality rates for stomach. (B) Male mortality rates for esophagus. (C) Female mortality rates for breast. (D) Female mortality rates for lung. The blue solid lines represent the mortality rates μ_i for Brazil, the orange solid lines represent the mortality rates μ_i for Ceará, and the orange dotted lines represent the effective mortality rates μ_i^{eff} for Ceará. We identified two key findings. For males, μ_i^{eff} for stomach and esophagus are higher and lower, respectively, than the corresponding μ_i in Ceará, indicating a strengthening of the pattern observed between Ceará and Brazil. For females, μ_i^{eff} for breast and lung are higher and lower, respectively, than the corresponding μ_i in Ceará, indicating a closer alignment with national patterns.

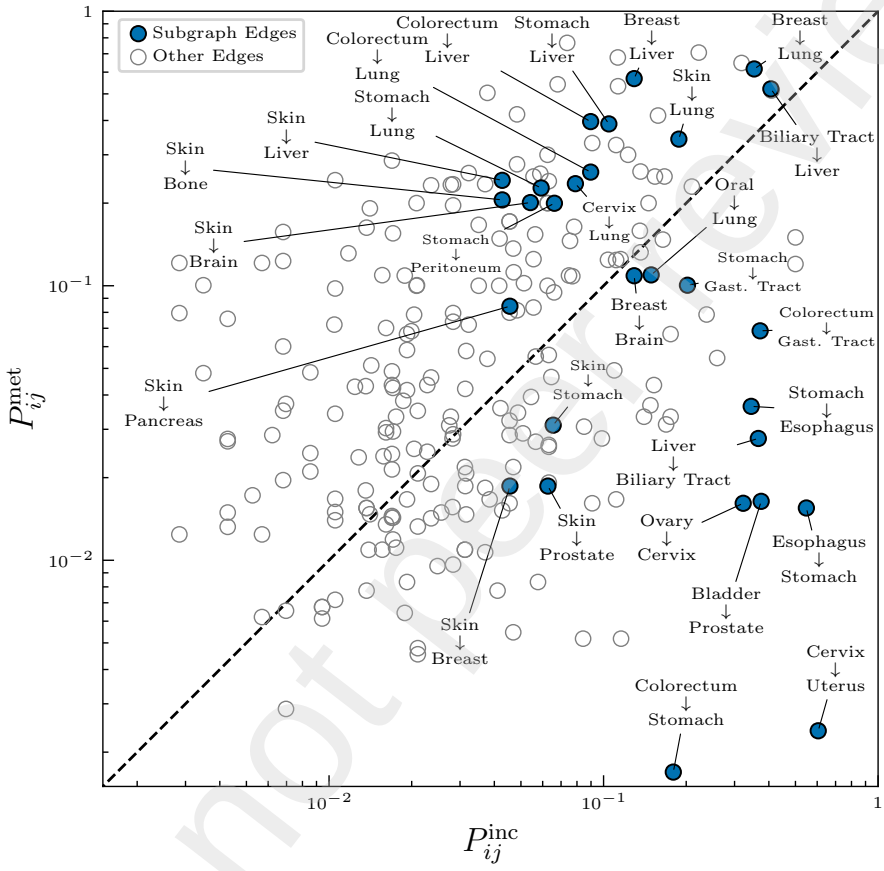

Figure 3: Scatter plot relating the inconsistency probability P_{ij}^{inc} to the metastasis probability P_{ij}^{met}. The plot shows the cancer site pairs in logarithmic scale. For each pair, P_{ij}^{inc} measures the probability that a cancer type in the Last Diagnosis (LD) is confused with another cancer type in the recorded Cause of Death (CoD), while P_{ij}^{met} measures the probability that a primary cancer type spreads to a metastatic site. Blue filled circles highlight pairs that belong to the core subnetwork, whereas empty circles represent all other edges. The dashed diagonal represents the equality line. Annotated arrows indicate notable pairs from the subnetwork. Among them, the Breast-to-Lung pair stands out by exhibiting high values for both probabilities, P_{ij}^{inc} and P_{ij}^{met}, suggesting that such inconsistencies may be associated with unreported metastatic deaths.

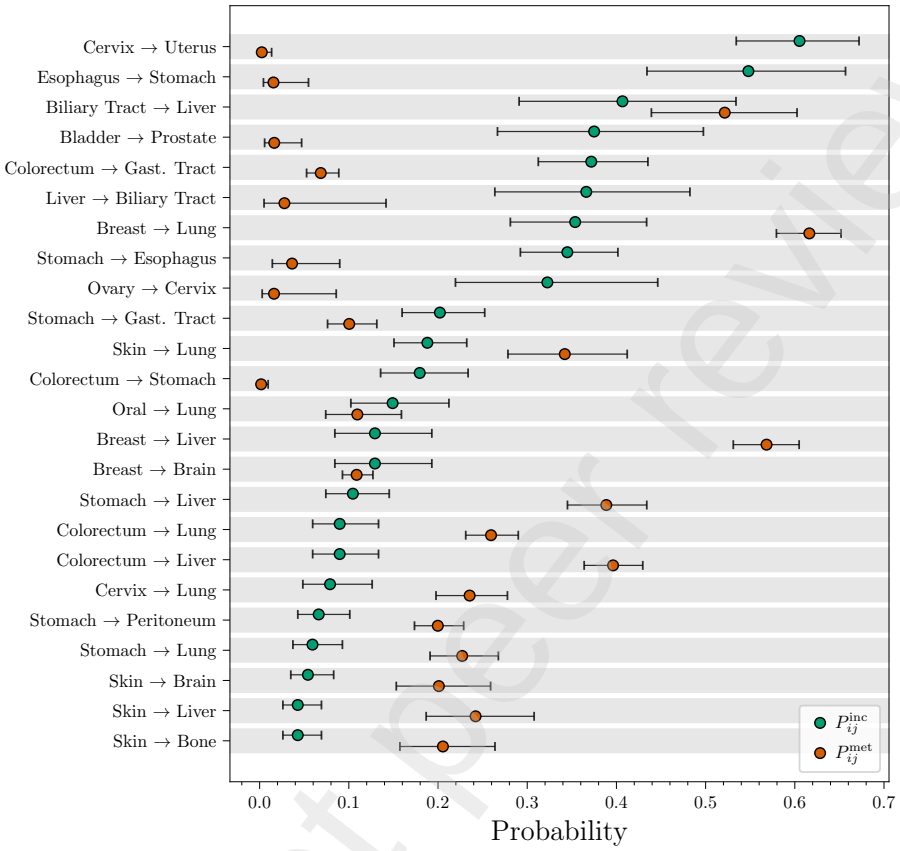

Figure 4: Inconsistency and metastatic probabilities for selected pairs of anatomical sites and theirs corresponding Wilson score Confidence Intervals, CI_{WS}. For each pair, the inconsistency probability P_{ij}^{inc} and the metastasis probability P_{ij}^{met} are shown as green and orange markers, respectively. The horizontal error bars represent the corresponding two-sided CI_{WS}. Notably, breast-to-lung pair exhibits high values for both probabilities along with narrow values of CI_{WS}, which may suggest that these inconsistencies are associated with unreported metastatic deaths.