Lung cancer survival dataset - Descriptive analysis

2024-10-22

Incidence - SEER Research Plus Specialized Data (with county in case listing), 17 Registries (excluding Alaska), November 2023 Submission (2000-2021)

Eligibility

- Known age > 18 y/o
- Site recode ICD-O-3 2023 Revision = Lung and bronchus
- Malignant behavior
- Baseline with no other cancers
- Exclude missing information

All the variables available

```
ls(mydata)
```

```
##   [1] "Adjusted.CS.site.specific.factor.7..2004.2017.varying.by.schema."
##   [2] "AFP.Post.Orchiectomy.Lab.Value.Recode..2010.."
##   [3] "AFP.Pretreatment.Interpretation.Recode..2010.."
##   [4] "Age.recode.with..1.year.olds"
##   [5] "Age.recode.with..1.year.olds.and.90."
##   [6] "Age.recode.with.single.ages.and.85."
##   [7] "Age.recode.with.single.ages.and.90."
##   [8] "AJCC.ID..2018.."
##   [9] "AJCC.stage.3rd.edition..1988.2003."
##  [10] "AYA.site.recode.2020.Revision"
##  [11] "B.Symptoms.Recode..2010.."
##  [12] "Behavior.code.ICD.O.3"
##  [13] "Behavior.recode.for.analysis"
##  [14] "Brain.Molecular.Markers..2018.."
##  [15] "Breast...Adjusted.AJCC.6th.M..1988.2015."
##  [16] "Breast...Adjusted.AJCC.6th.N..1988.2015."
##  [17] "Breast...Adjusted.AJCC.6th.Stage..1988.2015."
##  [18] "Breast...Adjusted.AJCC.6th.T..1988.2015."
##  [19] "Breast.Subtype..2010.."
##  [20] "Breslow.Thickness.Recode..2010.."
##  [21] "CA.125.Pretreatment.Interpretation.Recode..2010.."
##  [22] "CEA.Pretreatment.Interpretation.Recode..2010.."
##  [23] "Chemotherapy.recode..yes..no.unk."
##  [24] "Chromosome.19q..Loss.of.Heterozygosity..LOH..Recode..2010.."
##  [25] "Chromosome.1p..Loss.of.Heterozygosity..LOH..Recode..2010.."
##  [26] "CoC.Accredited.Flag..2018.."
```

```
##  [27] "COD.to.site.rec.KM"
##  [28] "COD.to.site.recode"
##  [29] "COD.to.site.recode.ICD.O.3.2023.Revision"
##  [30] "COD.to.site.recode.ICD.O.3.2023.Revision.Expanded..1999.."
##  [31] "Coding.system.EOD..1973.2003."
##  [32] "Combined.Summary.Stage..2004.."
##  [33] "CS.extension..2004.2015."
##  [34] "CS.lymph.nodes..2004.2015."
##  [35] "CS.mets.at.dx..2004.2015."
##  [36] "CS.Mets.Eval..2004.2015."
##  [37] "CS.Reg.Node.Eval..2004.2015."
##  [38] "CS.Schema...AJCC.6th.Edition"
##  [39] "CS.site.specific.factor.1..2004.2017.varying.by.schema."
##  [40] "CS.site.specific.factor.10..2004.2017.varying.by.schema."
##  [41] "CS.site.specific.factor.11..2004.2017.varying.by.schema."
##  [42] "CS.site.specific.factor.12..2004.2017.varying.by.schema."
##  [43] "CS.site.specific.factor.13..2004.2017.varying.by.schema."
##  [44] "CS.site.specific.factor.15..2004.2017.varying.by.schema."
##  [45] "CS.site.specific.factor.16..2004.2017.varying.by.schema."
##  [46] "CS.site.specific.factor.2..2004.2017.varying.by.schema."
##  [47] "CS.site.specific.factor.25..2004.2017.varying.by.schema."
##  [48] "CS.site.specific.factor.3..2004.2017.varying.by.schema."
##  [49] "CS.site.specific.factor.4..2004.2017.varying.by.schema."
##  [50] "CS.site.specific.factor.5..2004.2017.varying.by.schema."
##  [51] "CS.site.specific.factor.6..2004.2017.varying.by.schema."
##  [52] "CS.site.specific.factor.8..2004.2017.varying.by.schema."
##  [53] "CS.site.specific.factor.9..2004.2017.varying.by.schema."
##  [54] "CS.tumor.size..2004.2015."
##  [55] "CS.Tumor.Size.Ext.Eval..2004.2015."
##  [56] "CS.version.derived..2004.2015."
##  [57] "CS.version.input.current..2004.2015."
##  [58] "CS.version.input.original..2004.2015."
##  [59] "Cumulative.Expected..Calculated."
##  [60] "Derived.AJCC.M..6th.ed..2004.2015."
##  [61] "Derived.AJCC.M..7th.ed..2010.2015."
##  [62] "Derived.AJCC.N..6th.ed..2004.2015."
##  [63] "Derived.AJCC.N..7th.ed..2010.2015."
##  [64] "Derived.AJCC.Stage.Group..6th.ed..2004.2015."
##  [65] "Derived.AJCC.Stage.Group..7th.ed..2010.2015."
##  [66] "Derived.AJCC.T..6th.ed..2004.2015."
##  [67] "Derived.AJCC.T..7th.ed..2010.2015."
##  [68] "Derived.EOD.2018.M..2018.."
##  [69] "Derived.EOD.2018.N..2018.."
##  [70] "Derived.EOD.2018.Stage.Group..2018.."
##  [71] "Derived.EOD.2018.T..2018.."
##  [72] "Derived.HER2.Recode..2010.."
##  [73] "Derived.SEER.Cmb.Stg.Grp..2016.2017."
##  [74] "Derived.SEER.Combined.M..2016.2017."
##  [75] "Derived.SEER.Combined.M.Src..2016.2017."
##  [76] "Derived.SEER.Combined.N..2016.2017."
##  [77] "Derived.SEER.Combined.N.Src..2016.2017."
##  [78] "Derived.SEER.Combined.T..2016.2017."
##  [79] "Derived.SEER.Combined.T.Src..2016.2017."
##  [80] "Derived.Summary.Grade.2018..2018.."
```

```
##  [81] "Diagnostic.Confirmation"
##  [82] "End_date_field"
##  [83] "End.Calc.Vital.Status..Adjusted."
##  [84] "EOD.10...extent..1988.2003."
##  [85] "EOD.10...nodes..1988.2003."
##  [86] "EOD.10...Prostate.path.ext..1995.2003."
##  [87] "EOD.10...size..1988.2003."
##  [88] "EOD.4...extent..1983.1987."
##  [89] "EOD.4...nodes..1983.1987."
##  [90] "EOD.4...size..1983.1987."
##  [91] "EOD.Mets..2018.."
##  [92] "EOD.Primary.Tumor..2018.."
##  [93] "EOD.Regional.Nodes..2018.."
##  [94] "EOD.Schema.ID.Recode..2010.."
##  [95] "ER.Status.Recode.Breast.Cancer..1990.."
##  [96] "Expanded.EOD.1....CP53..1973.1982."
##  [97] "Expanded.EOD.1.2....CP53.54..1973.1982."
##  [98] "Expanded.EOD.10....CP62..1973.1982."
##  [99] "Expanded.EOD.11....CP63..1973.1982."
## [100] "Expanded.EOD.12....CP64..1973.1982."
## [101] "Expanded.EOD.13....CP65..1973.1982."
## [102] "Expanded.EOD.2....CP54..1973.1982."
## [103] "Expanded.EOD.3....CP55..1973.1982."
## [104] "Expanded.EOD.4....CP56..1973.1982."
## [105] "Expanded.EOD.5....CP57..1973.1982."
## [106] "Expanded.EOD.6....CP58..1973.1982."
## [107] "Expanded.EOD.7....CP59..1973.1982."
## [108] "Expanded.EOD.8....CP60..1973.1982."
## [109] "Expanded.EOD.9....CP61..1973.1982."
## [110] "Fibrosis.Score.Recode..2010.."
## [111] "Final.Interval.Expected..12.month."
## [112] "Final.Interval.Year..Calculated."
## [113] "FIPS_code"
## [114] "First.malignant.primary.indicator"
## [115] "Gestational.Trophoblastic.Prognostic.Scoring.Index.Recode..2010.."
## [116] "Gleason.Patterns.Clinical.Recode..2010.."
## [117] "Gleason.Patterns.Pathological.Recode..2010.."
## [118] "Gleason.Score.Clinical.Recode..2010.."
## [119] "Gleason.Score.Pathological.Recode..2010.."
## [120] "Grade.Clinical..2018.."
## [121] "Grade.Pathological..2018.."
## [122] "Grade.Recode..thru.2017."
## [123] "hCG.Post.Orchiectomy.Range.Recode..2010.."
## [124] "Histologic.Type.ICD.O.3"
## [125] "Histology.recode...broad.groupings"
## [126] "ICCC.site.recode.3rd.edition.IARC.2017"
## [127] "ICCC.site.recode.extended.3rd.edition.IARC.2017"
## [128] "ICD.O.3.Hist.behav"
## [129] "ICD.O.3.Hist.behav..malignant"
## [130] "ID"
## [131] "IHS.Link"
## [132] "Invasion.Beyond.Capsule.Recode..2010.."
## [133] "Ipsilateral.Adrenal.Gland.Involvement.Recode..2010.."
## [134] "Laterality"
```

```
## [135] "LDH.Post.Orchiectomy.Range.Recode..2010.."
## [136] "LDH.Pretreatment.Level.Recode..2010.."
## [137] "LN.Head.and.Neck.Levels.I.III.Recode..2010.."
## [138] "LN.Head.and.Neck.Levels.IV.V.Recode..2010.."
## [139] "LN.Head.and.Neck.Levels.VI.VII.Recode..2010.."
## [140] "LN.Head.and.Neck.Other.Recode..2010.."
## [141] "LN.Positive.Axillary.Level.I.II.Recode..2010.."
## [142] "Lymph.Node.Size.Recode..2010.."
## [143] "Lymph.vascular.Invasion..2004..varying.by.schema."
## [144] "Lymphoid.neoplasm.recode.2021.Revision"
## [145] "Lymphoma...Ann.Arbor.Stage..1983.2015."
## [146] "M.value...based.on.AJCC.3rd..1988.2003."
## [147] "Major.Vein.Involvement.Recode..2010.."
## [148] "Marital.status.at.diagnosis"
## [149] "Measured.Basal.Diameter.Recode..2010.."
## [150] "Measured.Thickness.Recode..2010.."
## [151] "Mets.at.DX.Distant.LN..2016.."
## [152] "Mets.at.DX.Other..2016.."
## [153] "Mitotic.Rate.Melanoma.Recode..2010.."
## [154] "N.value...based.on.AJCC.3rd..1988.2003."
## [155] "Number.of.Cores.Examined.Recode..2010.."
## [156] "Number.of.Cores.Positive.Recode..2010.."
## [157] "Number.of.Examined.Para.Aortic.Nodes.Recode..2010.."
## [158] "Number.of.Examined.Pelvic.Nodes.Recode..2010.."
## [159] "Number.of.Intervals..Calculated."
## [160] "Number.of.Positive.Para.Aortic.Nodes.Recode..2010.."
## [161] "Number.of.Positive.Pelvic.Nodes.Recode..2010.."
## [162] "Origin.recode.NHIA..Hispanic..Non.Hisp."
## [163] "Patient.ID"
## [164] "Perineural.Invasion.Recode..2010.."
## [165] "Peripheral.Blood.Involvement.Recode..2010.."
## [166] "Peritoneal.Cytology.Recode..2010.."
## [167] "Pleural.Effusion.Recode..2010.."
## [168] "PR.Status.Recode.Breast.Cancer..1990.."
## [169] "PRCDA.2020"
## [170] "PRCDA.Region"
## [171] "Primary.by.international.rules"
## [172] "Primary.Site"
## [173] "Primary.Site...labeled"
## [174] "Prostate.Pathological.Extension..2018.."
## [175] "PSA.Lab.Value.Recode..2010.."
## [176] "Race.and.origin.recode..NHW..NHB..NHAIAN..NHAPI..Hispanic."
## [177] "Race.ethnicity"
## [178] "Race.recode..W..B..AI..API."
## [179] "Race.recode..White..Black..Other."
## [180] "Radiation.recode"
## [181] "Radiation.to.Brain.or.CNS.Recode..1988.1997."
## [182] "Reason.no.cancer.directed.surgery"
## [183] "Record.number.recode"
## [184] "Regional.nodes.examined..1988.."
## [185] "Regional.nodes.positive..1988.."
## [186] "Residual.Tumor.Volume.Post.Cytoreduction.Recode..2010.."
## [187] "Response.to.Neoadjuvant.Therapy.Recode..2010.."
## [188] "RX.Summ..Reg.LN.Examined..1998.2002."
```

```
## [189] "RX.Summ..Scope.Reg.LN.Sur..2003.."
## [190] "RX.Summ..Surg.Oth.Reg.Dis..2003.."
## [191] "RX.Summ..Surg.Prim.Site..1998.."
## [192] "RX.Summ..Surg.Rad.Seq"
## [193] "RX.Summ..Systemic.Sur.Seq..2007.."
## [194] "Sarcomatoid.Features.Recode..2010.."
## [195] "Schema.ID..2018.."
## [196] "Scope.of.reg.lymph.nd.surg..1998.2002."
## [197] "SEER.Brain.and.CNS.Recode"
## [198] "SEER.cause.specific.death.classification"
## [199] "SEER.Combined.Mets.at.DX.bone..2010.."
## [200] "SEER.Combined.Mets.at.DX.brain..2010.."
## [201] "SEER.Combined.Mets.at.DX.liver..2010.."
## [202] "SEER.Combined.Mets.at.DX.lung..2010.."
## [203] "SEER.Combined.Summary.Stage.2000..2004.2017."
## [204] "SEER.historic.stage.A..1973.2015."
## [205] "SEER.modified.AJCC.stage.3rd..1988.2003."
## [206] "SEER.other.cause.of.death.classification"
## [207] "SEER.registry"
## [208] "SEER.registry..with.CA.and.GA.as.whole.states."
## [209] "Separate.Tumor.Nodules.Ipsilateral.Lung.Recode..2010.."
## [210] "Sequence.number"
## [211] "Sex"
## [212] "Site...mal.ins..least.detail."
## [213] "Site...mal.ins..mid.detail."
## [214] "Site...mal.ins..most.detail."
## [215] "Site...malignant..least.detail."
## [216] "Site...malignant..mid.detail."
## [217] "Site...malignant..most.detail."
## [218] "Site.recode...rare.tumors"
## [219] "Site.recode.ICD.O.3.2023.Revision"
## [220] "Site.recode.ICD.O.3.2023.Revision.Expanded"
## [221] "Site.recode.ICD.O.3.WHO.2008"
## [222] "Site.recode.ICD.O.3.WHO.2008..for.SIRs."
## [223] "Site.specific.surgery..1973.1997.varying.detail.by.year.and.site."
## [224] "SS.seq.....mal..least.detail."
## [225] "SS.seq.....mal..mid.detail."
## [226] "SS.seq.....mal..most.detail."
## [227] "SS.seq.....mal.ins..least.detail."
## [228] "SS.seq.....mal.ins..mid.detail."
## [229] "SS.seq.....mal.ins..most.detail."
## [230] "SS.seq...1975....mal..least.detail."
## [231] "SS.seq...1975....mal..mid.detail."
## [232] "SS.seq...1975....mal..most.detail."
## [233] "SS.seq...1975....mal.ins..least.detail."
## [234] "SS.seq...1975....mal.ins..mid.detail."
## [235] "SS.seq...1975....mal.ins..most.detail."
## [236] "SS.seq...1992....mal..least.detail."
## [237] "SS.seq...1992....mal..mid.detail."
## [238] "SS.seq...1992....mal..most.detail."
## [239] "SS.seq...1992....mal.ins..least.detail."
## [240] "SS.seq...1992....mal.ins..mid.detail."
## [241] "SS.seq...1992....mal.ins..most.detail."
## [242] "SS.seq...2000....mal..least.detail."
```

```
## [243] "SS.seq...2000....mal..mid.detail."
## [244] "SS.seq...2000....mal..most.detail."
## [245] "SS.seq...2000....mal.ins..least.detail."
## [246] "SS.seq...2000....mal.ins..mid.detail."
## [247] "SS.seq...2000....mal.ins..most.detail."
## [248] "Start_date_field"
## [249] "State.county"
## [250] "Summary.stage.2000..1998.2017."
## [251] "Surgery.of.oth.reg.dis.sites..1998.2002."
## [252] "Survival.months"
## [253] "Survival.months.flag"
## [254] "T.value...based.on.AJCC.3rd..1988.2003."
## [255] "Time.from.diagnosis.to.treatment.in.days.recode"
## [256] "TNM.7.CS.v0204..Schema..thru.2017."
## [257] "TNM.7.CS.v0204..Schema.recode"
## [258] "TNM.Edition.Number..2016.2017."
## [259] "Total.number.of.benign.borderline.tumors.for.patient"
## [260] "Total.number.of.in.situ.malignant.tumors.for.patient"
## [261] "Tumor.Deposits.Recode..2010.."
## [262] "Tumor.marker.1..1990.2003."
## [263] "Tumor.marker.2..1990.2003."
## [264] "Tumor.marker.3..1998.2003."
## [265] "Tumor.Size.Over.Time.Recode..1988.."
## [266] "Tumor.Size.Summary..2016.."
## [267] "Type.of.Reporting.Source"
## [268] "Ulceration.Recode..2010.."
## [269] "Visceral.and.Parietal.Pleural.Invasion.Recode..2010.."
## [270] "Vital.status.recode..study.cutoff.used."
## [271] "X..CRC.Test.Ever..age.50....sae.2004.2007."
## [272] "X..CRC.Test.Ever..age.50....sae.2008.2010."
## [273] "X..Current.Smoker..age.18....sae.1997.1999."
## [274] "X..Current.Smoker..age.18....sae.2000.2003."
## [275] "X..Current.Smoker..age.18....sae.2004.2007."
## [276] "X..Current.Smoker..age.18....sae.2008.2010."
## [277] "X..Current.Smoker..age.18....sae.2011.2013."
## [278] "X..Current.Smoker..age.18....sae.2014.2016."
## [279] "X..Current.Smoker..females.age.18....sae.1997.1999."
## [280] "X..Current.Smoker..females.age.18....sae.2000.2003."
## [281] "X..Current.Smoker..females.age.18....sae.2004.2007."
## [282] "X..Current.Smoker..females.age.18....sae.2008.2010."
## [283] "X..Current.Smoker..females.age.18....sae.2011.2013."
## [284] "X..Current.Smoker..females.age.18....sae.2014.2016."
## [285] "X..Current.Smoker..males.age.18....sae.1997.1999."
## [286] "X..Current.Smoker..males.age.18....sae.2000.2003."
## [287] "X..Current.Smoker..males.age.18....sae.2004.2007."
## [288] "X..Current.Smoker..males.age.18....sae.2008.2010."
## [289] "X..Current.Smoker..males.age.18....sae.2011.2013."
## [290] "X..Current.Smoker..males.age.18....sae.2014.2016."
## [291] "X..Endoscopy.Ever..age.50....sae.2004.2007."
## [292] "X..Endoscopy.Ever..age.50....sae.2008.2010."
## [293] "X..Ever.Smoker..age.18....sae.1997.1999."
## [294] "X..Ever.Smoker..age.18....sae.2000.2003."
## [295] "X..Ever.Smoker..age.18....sae.2004.2007."
## [296] "X..Ever.Smoker..age.18....sae.2008.2010."
```

```
## [297] "X..Ever.Smoker..females.age.18....sae.1997.1999."
## [298] "X..Ever.Smoker..females.age.18....sae.2000.2003."
## [299] "X..Ever.Smoker..females.age.18....sae.2004.2007."
## [300] "X..Ever.Smoker..females.age.18....sae.2008.2010."
## [301] "X..Ever.Smoker..males.age.18....sae.1997.1999."
## [302] "X..Ever.Smoker..males.age.18....sae.2000.2003."
## [303] "X..Ever.Smoker..males.age.18....sae.2004.2007."
## [304] "X..Ever.Smoker..males.age.18....sae.2008.2010."
## [305] "X..FOBT.Ever..age.50....sae.2004.2007."
## [306] "X..FOBT.Ever..age.50....sae.2008.2010."
## [307] "X..Former.Smoker..age.18....sae.2011.2013."
## [308] "X..Former.Smoker..age.18....sae.2014.2016."
## [309] "X..Former.Smoker..females.age.18....sae.2011.2013."
## [310] "X..Former.Smoker..females.age.18....sae.2014.2016."
## [311] "X..Former.Smoker..males.age.18....sae.2011.2013."
## [312] "X..Former.Smoker..males.age.18....sae.2014.2016."
## [313] "X..Long.term.former.smoker.quitting...1.year..age.18....sae.2011.2013."
## [314] "X..Long.term.former.smoker.quitting...1.year..age.18....sae.2014.2016."
## [315] "X..Long.term.former.smoker.quitting...1.year..females.age.18....sae.2011.2013."
## [316] "X..Long.term.former.smoker.quitting...1.year..females.age.18....sae.2014.2016."
## [317] "X..Long.term.former.smoker.quitting...1.year..males.age.18....sae.2011.2013."
## [318] "X..Long.term.former.smoker.quitting...1.year..males.age.18....sae.2014.2016."
## [319] "X..Mammography.within.2.years..age.40....sae.1997.1999."
## [320] "X..Mammography.within.2.years..age.40....sae.2000.2003."
## [321] "X..Mammography.within.2.years..age.40....sae.2004.2007."
## [322] "X..Mammography.within.2.years..age.40....sae.2008.2010."
## [323] "X..Mammography.within.2.years..age.40....sae.2011.2016."
## [324] "X..Mammography.within.2.years..age.50.74...sae.2011.2016."
## [325] "X..Pap.Smear.within.3.years..age.18....sae.1997.1999."
## [326] "X..Pap.Smear.within.3.years..age.18....sae.2000.2003."
## [327] "X..Pap.Smear.within.3.years..age.18....sae.2004.2007."
## [328] "X..Pap.Smear.within.3.years..age.18....sae.2008.2010."
## [329] "X..Pap.Smear.within.3.years..age.18....sae.2011.2016."
## [330] "X2.Digit.NS.EOD.part.1..1973.1982."
## [331] "X2.Digit.NS.EOD.part.2..1973.1982."
## [332] "X2.Digit.SS.EOD.part.1..1973.1982."
## [333] "X2.Digit.SS.EOD.part.2..1973.1982."
## [334] "X7th.Edition.Stage.Group.Recode..2016.2017."
## [335] "Year.of.death.recode"
## [336] "Year.of.diagnosis"
## [337] "Year.of.follow.up.recode"
```

Case numbers

```
nrow(mydata)
```

```
## [1] 703003
```

Variables pulled and prepped

```
mydata_1 <- mydata %>% select(Patient.ID, Sex, Year.of.diagnosis, Year.of.follow.up.recode, Year.of.dea

ls(mydata_1)
```

```
##  [1] "Age.recode.with.single.ages.and.90."
##  [2] "Chemotherapy.recode..yes..no.unk."
##  [3] "COD.to.site.recode.ICD.O.3.2023.Revision.Expanded..1999.."
##  [4] "Combined.Summary.Stage..2004.."
##  [5] "Derived.AJCC.Stage.Group..6th.ed..2004.2015."
##  [6] "Derived.EOD.2018.Stage.Group..2018.."
##  [7] "Derived.SEER.Cmb.Stg.Grp..2016.2017."
##  [8] "Histologic.Type.ICD.O.3"
##  [9] "Laterality"
## [10] "Marital.status.at.diagnosis"
## [11] "Patient.ID"
## [12] "Race.and.origin.recode..NHW..NHB..NHAIAN..NHAPI..Hispanic."
## [13] "Race.recode..White..Black..Other."
## [14] "Radiation.recode"
## [15] "Reason.no.cancer.directed.surgery"
## [16] "SEER.cause.specific.death.classification"
## [17] "SEER.modified.AJCC.stage.3rd..1988.2003."
## [18] "SEER.other.cause.of.death.classification"
## [19] "Sequence.number"
## [20] "Sex"
## [21] "Summary.stage.2000..1998.2017."
## [22] "Survival.months"
## [23] "Time.from.diagnosis.to.treatment.in.days.recode"
## [24] "Vital.status.recode..study.cutoff.used."
## [25] "Year.of.death.recode"
## [26] "Year.of.diagnosis"
## [27] "Year.of.follow.up.recode"
```

Age (Continuous)

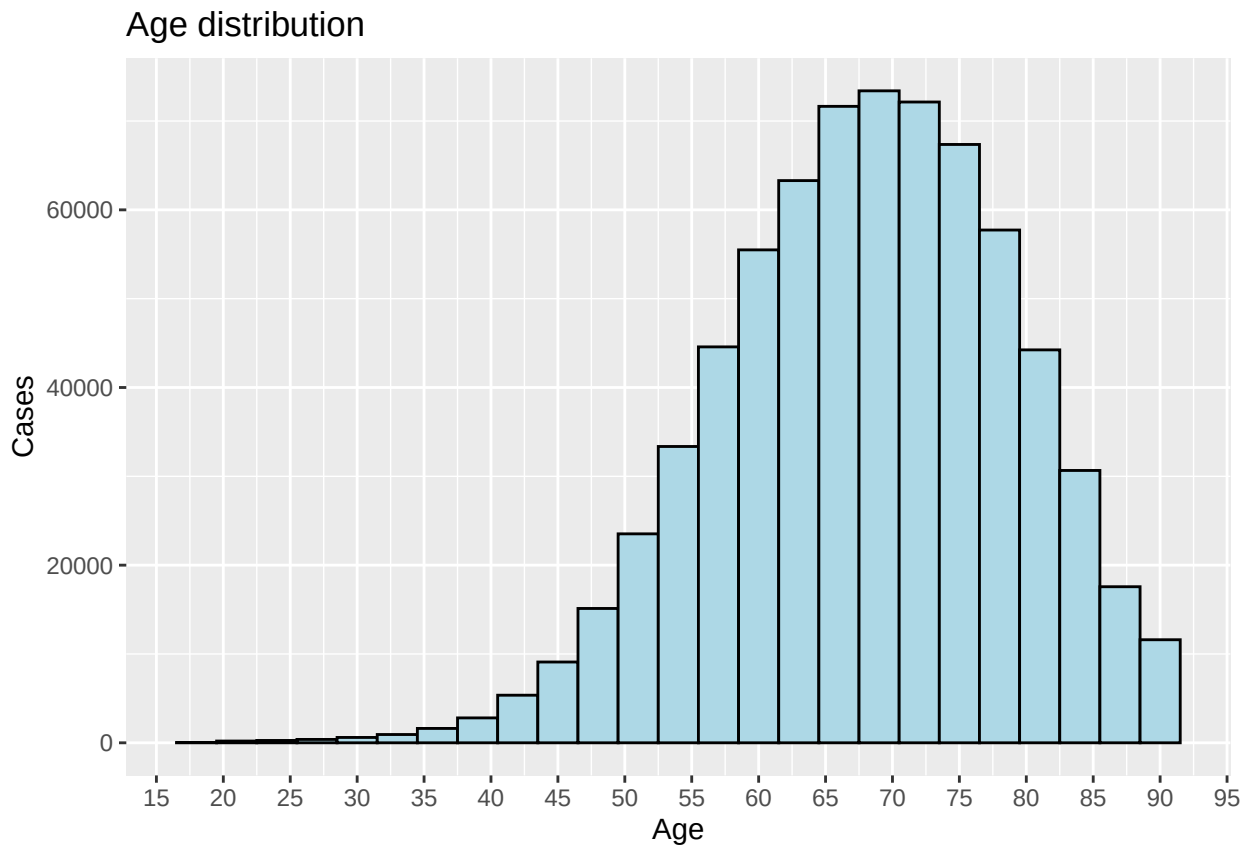
```
# Transform the "age" column to extract the numeric part
mydata_1 <- mydata_1 %>%
  mutate(age = as.numeric(gsub("[^0-9]", "", Age.recode.with.single.ages.and.90.)))

summary(mydata_1$age)  # Gives Min, 1st Qu., Median, Mean, 3rd Qu., Max
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19.00   61.00   68.00   67.89   76.00   90.00
```

```
sd(mydata_1$age)       # Standard Deviation
```

```
## [1] 10.89303
```

```r
# Create a histogram of the age column
ggplot(mydata_1, aes(x = age)) +
  geom_histogram(binwidth = 3, fill = "lightblue", color = "black") +
  scale_x_continuous(breaks = seq(10, 100, by = 5)) +
  xlab("Age") +
  ylab("Cases") +
  ggtitle("Age distribution")
```

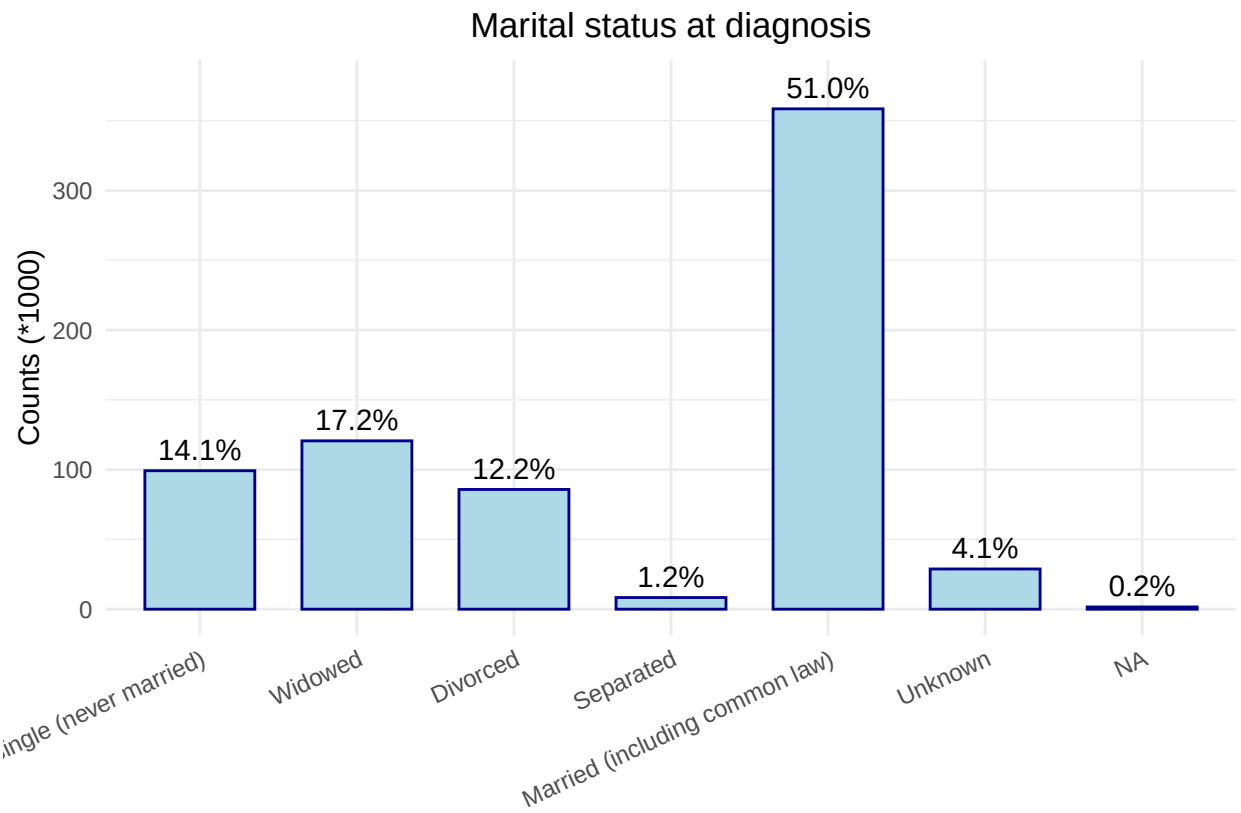


Marital status at diagnosis (Categorical)

```r
# Convert category to a factor and specify the order
mydata_1$Marital.status.at.diagnosis <- factor(mydata_1$Marital.status.at.diagnosis, levels = c("Single

# Create a summary table with counts and percentages
summary_table <- mydata_1 %>%
  group_by(Marital.status.at.diagnosis) %>%
  summarise(
    Count = n(),  # Count of each category
    Percentage = round((n() / nrow(mydata_1)) * 100, 1)  # Percentage of each category
  ) %>% ungroup()

# Display the summary table
kable(summary_table, caption = "")
```

Marital.status.at.diagnosis	Count	Percentage
Single (never married)	99231	14.1
Widowed	120649	17.2
Divorced	85811	12.2
Separated	8343	1.2
Married (including common law)	358558	51.0
Unknown	28815	4.1
NA	1596	0.2

```r
# Calculate percentage frequencies
mydata_1 <- mydata_1 %>%
  group_by(Marital.status.at.diagnosis) %>%
  mutate(Count = n()) %>%
  mutate(Percentage = Count / nrow(mydata_1) * 100) %>% ungroup()

# Create barplot
ggplot(mydata_1, aes(x = Marital.status.at.diagnosis)) +
  geom_bar(aes(y = after_stat(count) / 1000), width = 0.7, fill = "lightblue", color = "darkblue") +
  geom_text(aes(y = after_stat(count) / 1000, label = sprintf("%.1f%%", Percentage)),
            stat = "count", vjust = -0.5, color = "black") +  # Add percentages on top of bars
  labs(title = "Marital status at diagnosis", x = "", y = "Counts (*1000)") +
  scale_y_continuous(limits = c(0, 375)) +
  theme_minimal() +  # Optional: Use a minimal theme
  theme(
    axis.text.x = element_text(angle = 25, hjust = 1)  # Angle x-axis labels
  ) +
  theme(plot.title = element_text(hjust = 0.5))  # Center the title
```

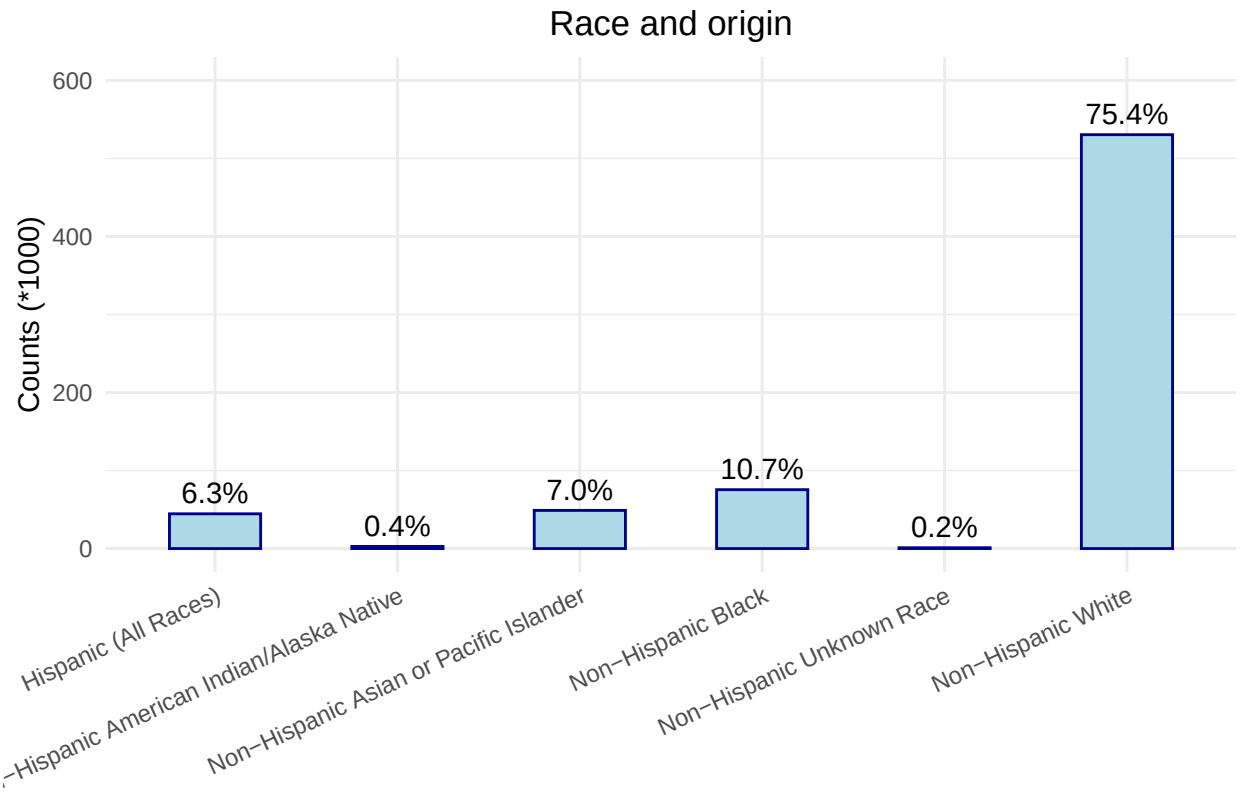


Race and origin 1 (Categorical)

```
# Create a summary table with counts and percentages
summary_table <- mydata_1 %>%
  group_by(Race.and.origin.recode..NHW..NHB..NHAIAN..NHAPI..Hispanic.) %>%
  summarise(
    Count = n(),  # Count of each category
    Percentage = round((n() / nrow(mydata_1)) * 100, 1)  # Percentage of each category
  ) %>% ungroup()

# Display the summary table
kable(summary_table, caption = "")
```

Race.and.origin.recode..NHW..NHB..NHAIAN..NHAPI..Hispanic.	Count	Percentage
Hispanic (All Races)	44516	6.3
Non-Hispanic American Indian/Alaska Native	2707	0.4
Non-Hispanic Asian or Pacific Islander	48884	7.0
Non-Hispanic Black	75449	10.7
Non-Hispanic Unknown Race	1076	0.2
Non-Hispanic White	530371	75.4

```r
# Calculate percentage frequencies
mydata_1 <- mydata_1 %>%
  group_by(Race.and.origin.recode..NHW..NHB..NHAIAN..NHAPI..Hispanic.) %>%
  mutate(Count = n()) %>%
  mutate(Percentage = Count / nrow(mydata_1) * 100) %>% ungroup()

# Create barplot
ggplot(mydata_1, aes(x = Race.and.origin.recode..NHW..NHB..NHAIAN..NHAPI..Hispanic.)) +
  geom_bar(aes(y = after_stat(count) / 1000), width = 0.5, fill = "lightblue", color = "darkblue") +
  geom_text(aes(y = after_stat(count) / 1000, label = sprintf("%.1f%%", Percentage)),
            stat = "count", vjust = -0.5, color = "black") +  # Add percentages on top of bars
  labs(title = "Race and origin", x = "", y = "Counts (*1000)") +
  scale_y_continuous(limits = c(0, 600)) +
  theme_minimal() +  # Optional: Use a minimal theme
  theme(
    axis.text.x = element_text(angle = 25, hjust = 1)  # Angle x-axis labels
  ) +
  theme(plot.title = element_text(hjust = 0.5))  # Center the title
```



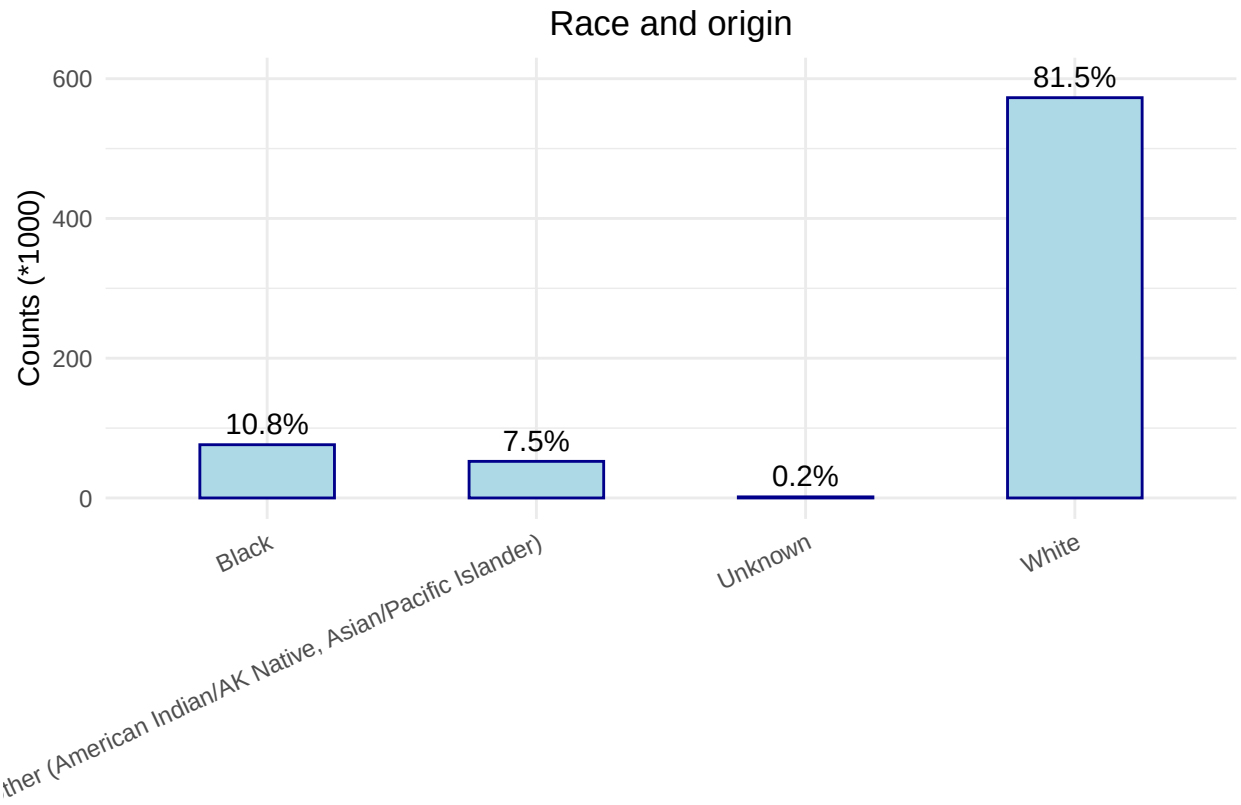
Race and origin 2 (Categorical)

```r
# Create a summary table with counts and percentages
summary_table <- mydata_1 %>%
```

```r
  group_by(Race.recode..White..Black..Other.) %>%
  summarise(
    Count = n(),  # Count of each category
    Percentage = round((n() / nrow(mydata_1)) * 100, 1)  # Percentage of each category
  ) %>% ungroup()

# Display the summary table
kable(summary_table, caption = "")
```

Race.recode..White..Black..Other.	Count	Percentage
Black	76221	10.8
Other (American Indian/AK Native, Asian/Pacific Islander)	52391	7.5
Unknown	1566	0.2
White	572825	81.5

```r
# Calculate percentage frequencies
mydata_1 <- mydata_1 %>%
  group_by(Race.recode..White..Black..Other.) %>%
  mutate(Count = n()) %>%
  mutate(Percentage = Count / nrow(mydata_1) * 100) %>% ungroup()

# Create barplot
ggplot(mydata_1, aes(x = Race.recode..White..Black..Other.)) +
  geom_bar(aes(y = after_stat(count) / 1000), width = 0.5, fill = "lightblue", color = "darkblue") +
  geom_text(aes(y = after_stat(count) / 1000, label = sprintf("%.1f%%", Percentage)),
            stat = "count", vjust = -0.5, color = "black") +  # Add percentages on top of bars
  labs(title = "Race and origin", x = "", y = "Counts (*1000)") +
  scale_y_continuous(limits = c(0, 600)) +
  theme_minimal() +  # Optional: Use a minimal theme
  theme(
    axis.text.x = element_text(angle = 25, hjust = 1)  # Angle x-axis labels
  ) +
  theme(plot.title = element_text(hjust = 0.5))  # Center the title
```



Histological type

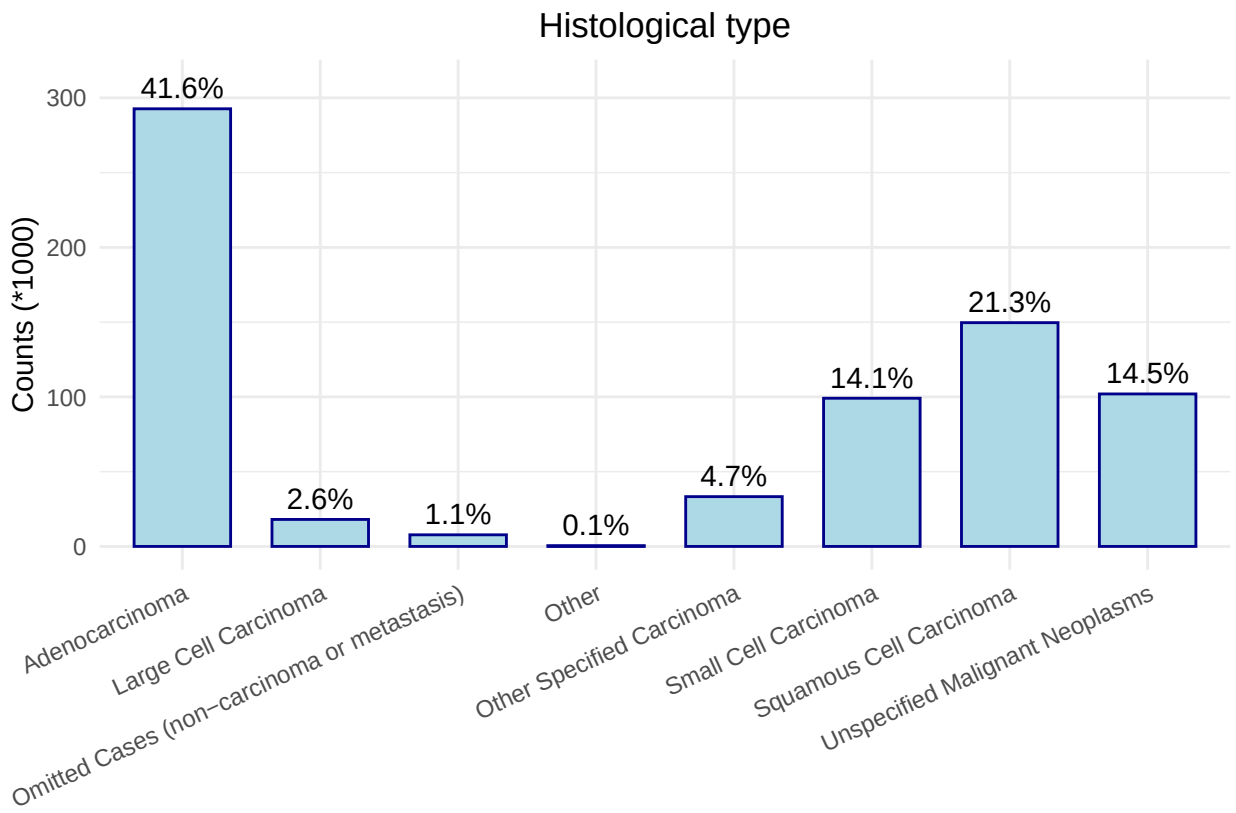
Categorized based on ICD code: Lewis DR, Check DP, Caporaso NE, Travis WD, Devesa SS. US lung cancer trends by histologic type. Cancer. 2014 Sep 15;120(18):2883-92. doi: 10.1002/cncr.28749. Epub 2014 Aug 11. PMID: 25113306; PMCID: PMC4187244.

```
#Categorical variable
mydata_1 <- mydata_1 %>%
  mutate(histological_type = case_when(
    Histologic.Type.ICD.O.3 %in% c(8051:8052, 8070:8076, 8078, 8083:8084, 8090, 8094, 8120, 8123) ~ "Squ
    Histologic.Type.ICD.O.3 %in% c(8002, 8041:8045) ~ "Small Cell Carcinoma",
    Histologic.Type.ICD.O.3 %in% c(8015, 8050, 8140:8141, 8143:8145, 8147, 8190, 8201, 8211,
                                   8250:8255, 8260, 8290, 8310, 8320, 8323, 8333, 8401,
                                   8440, 8470:8471, 8480:8481, 8490, 8503, 8507, 8550,
                                   8570:8572, 8574, 8576) ~ "Adenocarcinoma",
    Histologic.Type.ICD.O.3 %in% c(8012:8014, 8021, 8034, 8082) ~ "Large Cell Carcinoma",
    Histologic.Type.ICD.O.3 %in% c(8003:8004, 8022, 8030:8035, 8200, 8240:8241, 8243:8246,
                                   8249, 8430, 8525, 8560, 8562, 8575) ~ "Other Specified Carcinoma",
    Histologic.Type.ICD.O.3 %in% c(8010:8011, 8020, 8230, 8046, 8000:8001) ~ "Unspecified Malignant Neop
    Histologic.Type.ICD.O.3 %in% c(8580:9999, 8005, 8095, 8124, 8130, 8146, 8160, 8170,
                                   8231, 8247, 8263, 8312, 8340:8341, 8350, 8370, 8441,
                                   8460, 8500, 8501, 8510, 8524, 8530, 8551) ~ "Omitted Cases (non-carci
    TRUE ~ "Other"  # Default case for any unmatched codes
  ))
```

```r
# Create a summary table with counts and percentages
summary_table <- mydata_1 %>%
  group_by(histological_type) %>%
  summarise(
    Count = n(),  # Count of each category
    Percentage = round((n() / nrow(mydata_1)) * 100, 1)  # Percentage of each category
  ) %>% ungroup()

# Display the summary table
kable(summary_table, caption = "")
```

histological_type	Count	Percentage
Adenocarcinoma	292676	41.6
Large Cell Carcinoma	18040	2.6
Omitted Cases (non-carcinoma or metastasis)	7753	1.1
Other	514	0.1
Other Specified Carcinoma	33298	4.7
Small Cell Carcinoma	99104	14.1
Squamous Cell Carcinoma	149641	21.3
Unspecified Malignant Neoplasms	101977	14.5

```r
# Calculate percentage frequencies
mydata_1 <- mydata_1 %>%
  group_by(histological_type) %>%
  mutate(Count = n()) %>%
  mutate(Percentage = Count / nrow(mydata_1) * 100) %>% ungroup()

# Create barplot
ggplot(mydata_1, aes(x = histological_type)) +
  geom_bar(aes(y = after_stat(count) / 1000), width = 0.7, fill = "lightblue", color = "darkblue") +
  geom_text(aes(y = after_stat(count) / 1000, label = sprintf("%.1f%%", Percentage)),
            stat = "count", vjust = -0.5, color = "black") +  # Add percentages on top of bars
  labs(title = "Histological type", y = "Counts (*1000)", x = "") +
  scale_y_continuous(limits = c(0, 310)) +
  theme_minimal() +  # Optional: Use a minimal theme
  theme(
    axis.text.x = element_text(angle = 25, hjust = 1)  # Angle x-axis labels
  ) +
  theme(plot.title = element_text(hjust = 0.5))  # Center the title
```

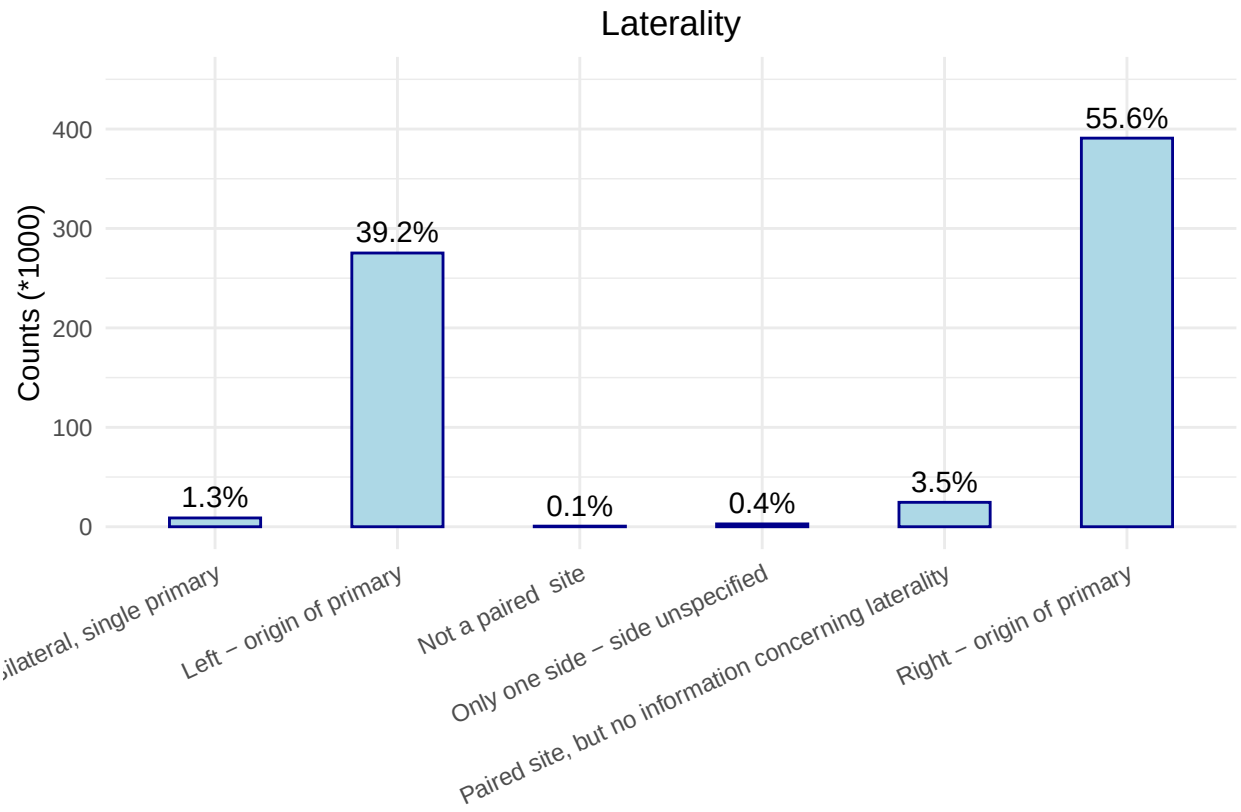


Laterality (Categorical)

```
# Create a summary table with counts and percentages
summary_table <- mydata_1 %>%
  group_by(Laterality) %>%
  summarise(
    Count = n(),  # Count of each category
    Percentage = round((n() / nrow(mydata_1)) * 100, 1)  # Percentage of each category
  ) %>% ungroup()

# Display the summary table
kable(summary_table, caption = "")
```

Laterality	Count	Percentage
Bilateral, single primary	8843	1.3
Left - origin of primary	275296	39.2
Not a paired site	605	0.1
Only one side - side unspecified	2838	0.4
Paired site, but no information concerning laterality	24579	3.5
Right - origin of primary	390842	55.6

```r
# Calculate percentage frequencies
mydata_1 <- mydata_1 %>%
  group_by(Laterality) %>%
  mutate(Count = n()) %>%
  mutate(Percentage = Count / nrow(mydata_1) * 100) %>% ungroup()

# Create barplot
ggplot(mydata_1, aes(x = Laterality)) +
  geom_bar(aes(y = after_stat(count) / 1000), width = 0.5, fill = "lightblue", color = "darkblue") +
  geom_text(aes(y = after_stat(count) / 1000, label = sprintf("%.1f%%", Percentage)),
            stat = "count", vjust = -0.5, color = "black") +  # Add percentages on top of bars
  labs(title = "Laterality", x = "", y = "Counts (*1000)") +
  scale_y_continuous(limits = c(0, 450)) +
  theme_minimal() +  # Optional: Use a minimal theme
  theme(
    axis.text.x = element_text(angle = 25, hjust = 1)  # Angle x-axis labels
  ) +
  theme(plot.title = element_text(hjust = 0.5))  # Center the title
```



Clinical stage (Categorical)

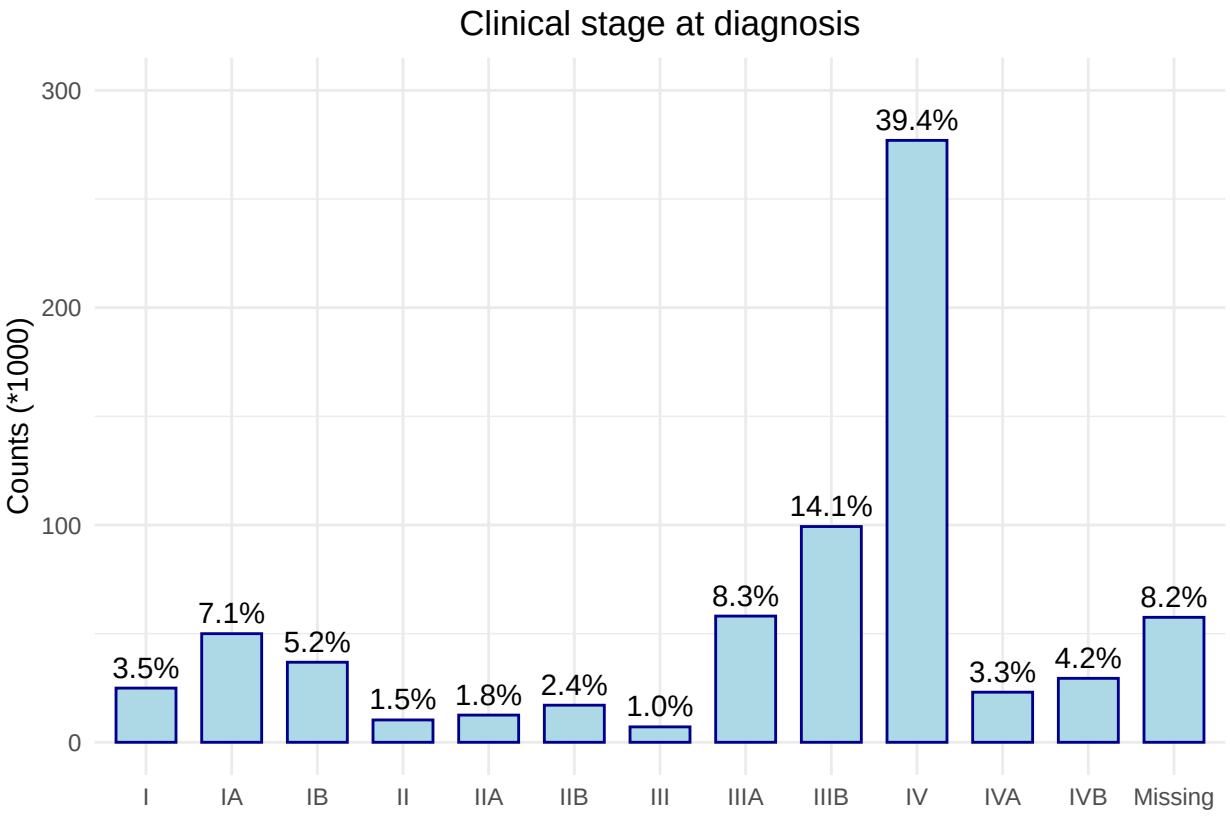
```r
stage00_03 <- mydata_1 %>% filter(Year.of.diagnosis %in% c(2000:2003)) %>%
  mutate(stage = case_when(
```

```
    SEER.modified.AJCC.stage.3rd..1988.2003. == 10 ~ "I",
    SEER.modified.AJCC.stage.3rd..1988.2003. == 20 ~ "II",
    SEER.modified.AJCC.stage.3rd..1988.2003. == 31 ~ "IIIA",
    SEER.modified.AJCC.stage.3rd..1988.2003. == 32 ~ "IIIB",
    SEER.modified.AJCC.stage.3rd..1988.2003. == 40 ~ "IV",
    TRUE ~ "Missing"  # Default case for any unmatched codes
  )) %>% select(stage, Patient.ID)

stage04_15 <- mydata_1 %>% filter(Year.of.diagnosis %in% c(2004:2015)) %>%
  mutate(stage = case_when(
    Derived.AJCC.Stage.Group..6th.ed..2004.2015. == "IA" ~ "IA",
    Derived.AJCC.Stage.Group..6th.ed..2004.2015. == "IB" ~ "IB",
    Derived.AJCC.Stage.Group..6th.ed..2004.2015. == "IIA" ~ "IIA",
    Derived.AJCC.Stage.Group..6th.ed..2004.2015. == "IIB" ~ "IIB",
    Derived.AJCC.Stage.Group..6th.ed..2004.2015. == "IIIA" ~ "IIIA",
    Derived.AJCC.Stage.Group..6th.ed..2004.2015. == "IIIB" ~ "IIIB",
    Derived.AJCC.Stage.Group..6th.ed..2004.2015. == "IV" ~ "IV",
    TRUE ~ "Missing"  # Default case for any unmatched codes
  )) %>% select(stage, Patient.ID)

stage16_17 <- mydata_1 %>% filter(Year.of.diagnosis %in% c(2016:2017)) %>%
  mutate(stage = case_when(
    Derived.SEER.Cmb.Stg.Grp..2016.2017. == "0" ~ "I",
    Derived.SEER.Cmb.Stg.Grp..2016.2017. == "1A" ~ "IA",
    Derived.SEER.Cmb.Stg.Grp..2016.2017. == "1B" ~ "IB",
    Derived.SEER.Cmb.Stg.Grp..2016.2017. == "2" ~ "II",
    Derived.SEER.Cmb.Stg.Grp..2016.2017. == "2A" ~ "IIA",
    Derived.SEER.Cmb.Stg.Grp..2016.2017. == "2B" ~ "IIB",
    Derived.SEER.Cmb.Stg.Grp..2016.2017. == "3" ~ "III",
    Derived.SEER.Cmb.Stg.Grp..2016.2017. == "3A" ~ "IIIA",
    Derived.SEER.Cmb.Stg.Grp..2016.2017. == "3B" ~ "IIIB",
    Derived.SEER.Cmb.Stg.Grp..2016.2017. == "4" ~ "IV",
    TRUE ~ "Missing"  # Default case for any unmatched codes
  )) %>% select(stage, Patient.ID)

stage18_21 <- mydata_1 %>% filter(Year.of.diagnosis %in% c(2018:2021)) %>%
  mutate(stage = case_when(
    Derived.EOD.2018.Stage.Group..2018.. == "0" ~ "I",
    Derived.EOD.2018.Stage.Group..2018.. == "1A1" ~ "IA",
    Derived.EOD.2018.Stage.Group..2018.. == "1B2" ~ "IB",
    Derived.EOD.2018.Stage.Group..2018.. == "1A3" ~ "II",
    Derived.EOD.2018.Stage.Group..2018.. == "1B" ~ "IIA",
    Derived.EOD.2018.Stage.Group..2018.. == "2A" ~ "IIB",
    Derived.EOD.2018.Stage.Group..2018.. == "2B" ~ "III",
    Derived.EOD.2018.Stage.Group..2018.. == "3" ~ "IIIA",
    Derived.EOD.2018.Stage.Group..2018.. == "3A" ~ "IIIB",
    Derived.EOD.2018.Stage.Group..2018.. == "3B" ~ "IV",
    Derived.EOD.2018.Stage.Group..2018.. == "3C" ~ "IV",
    Derived.EOD.2018.Stage.Group..2018.. == "4" ~ "IV",
    Derived.EOD.2018.Stage.Group..2018.. == "4A" ~ "IVA",
    Derived.EOD.2018.Stage.Group..2018.. == "4B" ~ "IVB",
    TRUE ~ "Missing"  # Default case for any unmatched codes
  )) %>% select(stage, Patient.ID)
```

```r
stage_joined_data <- stage00_03 %>% full_join(stage04_15, by = "Patient.ID") %>%
  full_join(stage16_17, by = "Patient.ID") %>%
  full_join(stage18_21, by = "Patient.ID") %>% mutate(final_stage = case_when(
    !is.na(stage.x) & is.na(stage.y) & is.na(stage.x.x) & is.na(stage.y.y) ~ stage.x,
    is.na(stage.x) & !is.na(stage.y) & is.na(stage.x.x) & is.na(stage.y.y) ~ stage.y,
    is.na(stage.x) & is.na(stage.y) & !is.na(stage.x.x) & is.na(stage.y.y) ~ stage.x.x,
    is.na(stage.x) & is.na(stage.y) & is.na(stage.x.x) & !is.na(stage.y.y) ~ stage.y.y,
    is.na(stage.x) & is.na(stage.y) & is.na(stage.x.x) & is.na(stage.y.y) ~ NA_character_,
  )) %>% select(final_stage, Patient.ID)

# Create a summary table with counts and percentages
summary_table <- stage_joined_data %>%
  group_by(final_stage) %>%
  summarise(
    Count = n(),  # Count of each category
    Percentage = round((n() / nrow(stage_joined_data)) * 100, 1)  # Percentage of each category
  ) %>% ungroup()

# Display the summary table
kable(summary_table, caption = "")
```

final_stage	Count	Percentage
I	24928	3.5
IA	49979	7.1
IB	36836	5.2
II	10277	1.5
IIA	12511	1.8
IIB	17067	2.4
III	7088	1.0
IIIA	58079	8.3
IIIB	99279	14.1
IV	276974	39.4
IVA	23059	3.3
IVB	29428	4.2
Missing	57498	8.2

```r
# Calculate percentage frequencies
stage_joined_data <- stage_joined_data %>%
  group_by(final_stage) %>%
  mutate(Count = n()) %>%
  mutate(Percentage = Count / nrow(stage_joined_data) * 100) %>% ungroup()

# Create barplot
ggplot(stage_joined_data, aes(x = final_stage)) +
  geom_bar(aes(y = after_stat(count) / 1000), width = 0.7, fill = "lightblue", color = "darkblue") +
  geom_text(aes(y = after_stat(count) / 1000, label = sprintf("%.1f%%", Percentage)),
            stat = "count", vjust = -0.5, color = "black") +  # Add percentages on top of bars
  labs(title = "Clinical stage at diagnosis", y = "Counts (*1000)", x = "") +
  scale_y_continuous(limits = c(0, 300)) +
  theme_minimal() +  # Optional: Use a minimal theme
  theme(plot.title = element_text(hjust = 0.5))  # Center the title
```



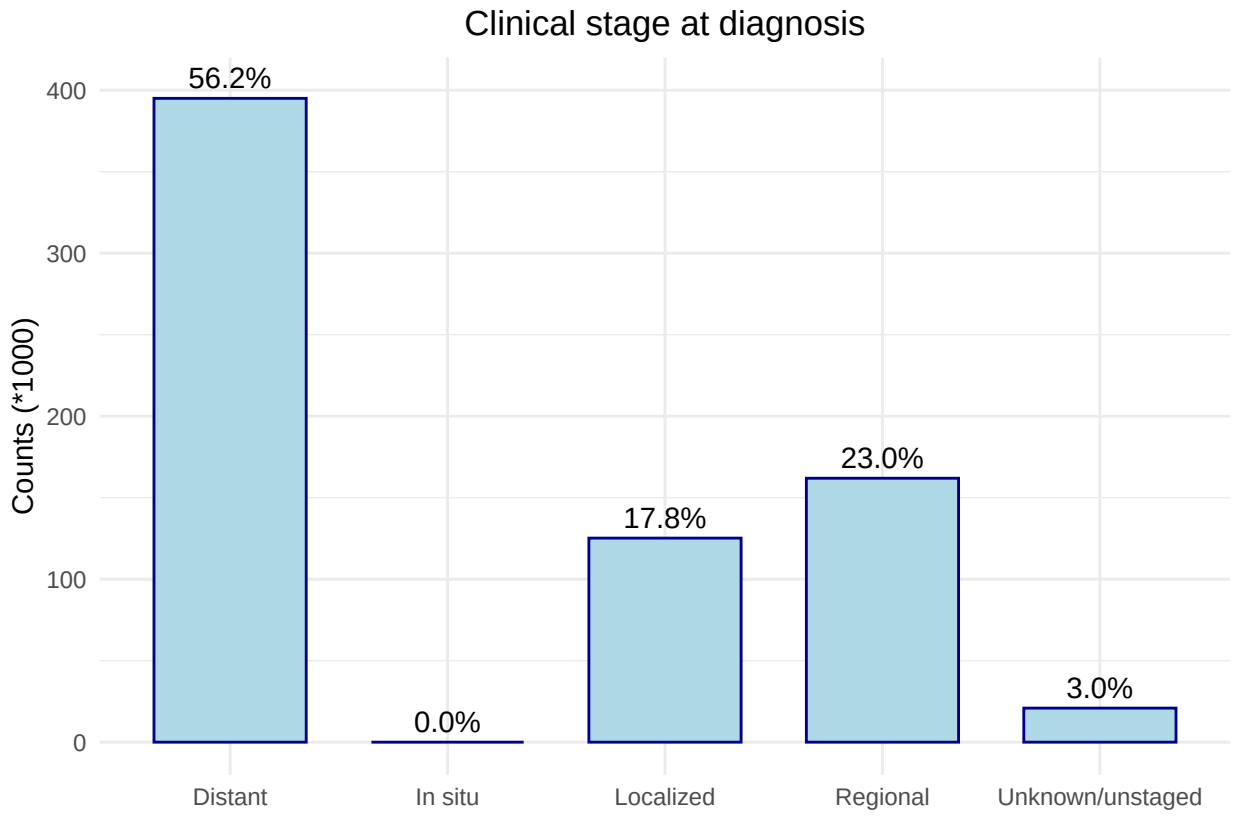
Clinical stage summary (Categorical)

```
mydata_1 <- mydata_1 %>%
  mutate(combined_stage = case_when(
    Year.of.diagnosis < 2004 ~ Summary.stage.2000..1998.2017.,  # Use stage_A if year <= 2004
    Year.of.diagnosis >= 2004 ~ Combined.Summary.Stage..2004..    # Use stage_B if year > 2004
  ))

# Create a summary table with counts and percentages
summary_table <- mydata_1 %>%
  group_by(combined_stage) %>%
  summarise(
    Count = n(),  # Count of each category
    Percentage = round((n() / nrow(mydata_1)) * 100, 1)  # Percentage of each category
  ) %>% ungroup()

# Display the summary table
kable(summary_table, caption = "")
```

combined_stage	Count	Percentage
Distant	394995	56.2
In situ	3	0.0

combined_stage	Count	Percentage
Localized	125192	17.8
Regional	161925	23.0
Unknown/unstaged	20888	3.0

```r
# Calculate percentage frequencies
mydata_1 <- mydata_1 %>%
  group_by(combined_stage) %>%
  mutate(Count = n()) %>%
  mutate(Percentage = Count / nrow(mydata_1) * 100) %>% ungroup()

# Create barplot
ggplot(mydata_1, aes(x = combined_stage)) +
  geom_bar(aes(y = after_stat(count) / 1000), width = 0.7, fill = "lightblue", color = "darkblue") +
  geom_text(aes(y = after_stat(count) / 1000, label = sprintf("%.1f%%", Percentage)),
            stat = "count", vjust = -0.5, color = "black") +  # Add percentages on top of bars
  labs(title = "Clinical stage at diagnosis", y = "Counts (*1000)", x = "") +
  scale_y_continuous(limits = c(0, 400)) +
  theme_minimal() +  # Optional: Use a minimal theme
  theme(plot.title = element_text(hjust = 0.5))  # Center the title
```

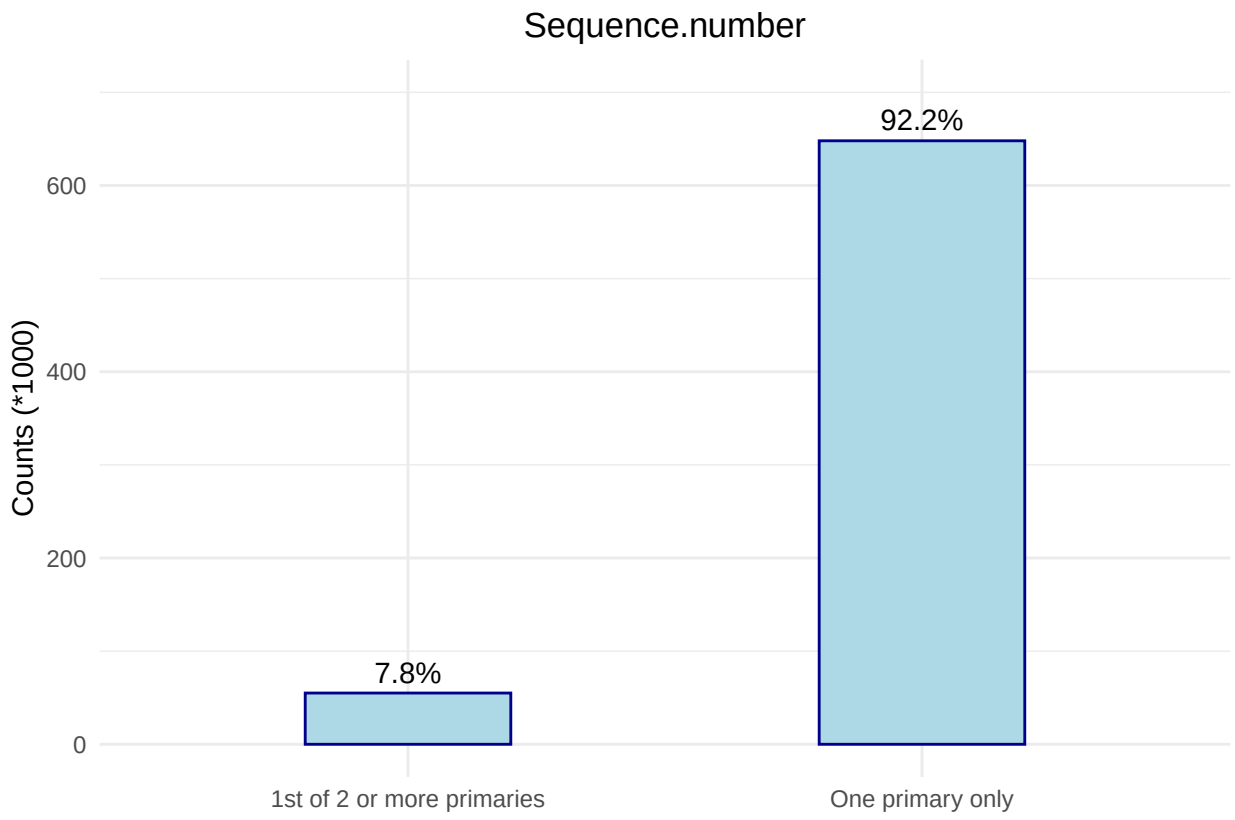


Tumor sequence number

```r
# Create a summary table with counts and percentages
summary_table <- mydata_1 %>%
  group_by(Sequence.number) %>%
  summarise(
    Count = n(),  # Count of each category
    Percentage = round((n() / nrow(mydata_1)) * 100, 1)  # Percentage of each category
  ) %>% ungroup()

# Display the summary table
kable(summary_table, caption = "")
```

Sequence.number	Count	Percentage
1st of 2 or more primaries	55034	7.8
One primary only	647969	92.2

```r
# Calculate percentage frequencies
mydata_1 <- mydata_1 %>%
  group_by(Sequence.number) %>%
  mutate(Count = n()) %>%
  mutate(Percentage = Count / nrow(mydata_1) * 100) %>% ungroup()

# Create barplot
ggplot(mydata_1, aes(x = Sequence.number)) +
  geom_bar(aes(y = after_stat(count) / 1000), width = 0.4, fill = "lightblue", color = "darkblue") +
  geom_text(aes(y = after_stat(count) / 1000, label = sprintf("%.1f%%", Percentage)),
            stat = "count", vjust = -0.5, color = "black") +  # Add percentages on top of bars
  labs(title = "Sequence.number", y = "Counts (*1000)", x = "") +
  scale_y_continuous(limits = c(0, 700)) +
  theme_minimal() +  # Optional: Use a minimal theme
  theme(plot.title = element_text(hjust = 0.5))  # Center the title
```



Days from diagnosis to treatment

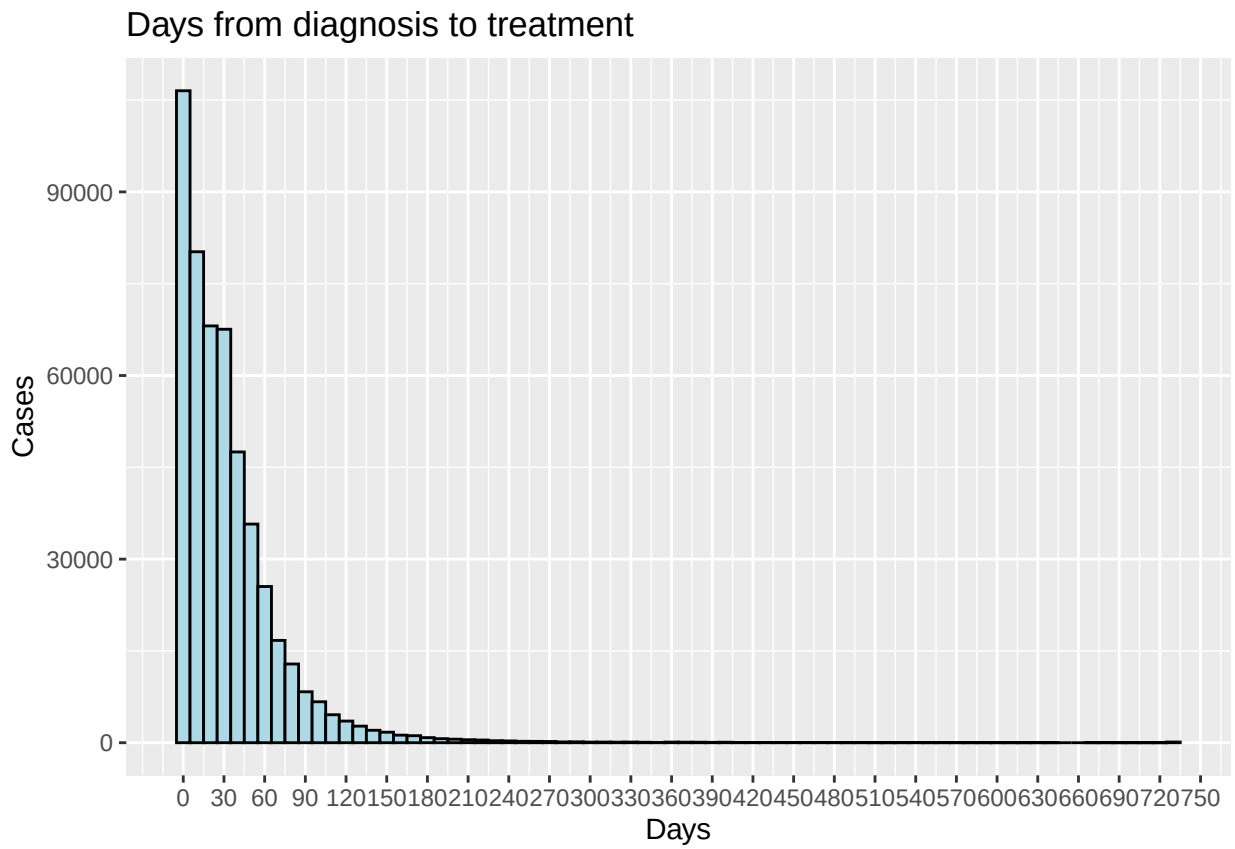
```
mydata_1 <- mydata_1 %>%
  mutate(time_from_diagnosis_to_treatment = ifelse(Time.from.diagnosis.to.treatment.in.days.recode == "
  mutate(time_from_diagnosis_to_treatment_1 = ifelse(time_from_diagnosis_to_treatment == "Unable to cal
  mutate(time_from_diagnosis_to_treatment_2 = as.numeric(time_from_diagnosis_to_treatment_1)) %>%
  mutate(radiation = case_when(
    Radiation.recode %in% c("Beam radiation", "Combination of beam with implants or isotopes", "Radiati
    Radiation.recode %in% c("None/Unknown", "Recommended, unknown if administered", "Refused (1988+)") 
  )) %>%
  mutate(surgery = case_when(
    Reason.no.cancer.directed.surgery == "Surgery performed" ~ "Yes",
    Reason.no.cancer.directed.surgery %in% c("Not performed, patient died prior to recommended surgery"
  ))

summary(mydata_1$time_from_diagnosis_to_treatment_2)  # Gives Min, 1st Qu., Median, Mean, 3rd Qu., Max
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00    7.00   25.00   34.31   47.00  731.00  204242
```

```
# Create a histogram of the age column
ggplot(mydata_1, aes(x = time_from_diagnosis_to_treatment_2)) +
  geom_histogram(binwidth = 10, fill = "lightblue", color = "black") +
```

```
  scale_x_continuous(breaks = seq(0, 750, by = 30)) +
  xlab("Days") +
  ylab("Cases") +
  ggtitle("Days from diagnosis to treatment")
```

```
## Warning: Removed 204242 rows containing non-finite outside the scale range
## (‘stat_bin()‘).
```

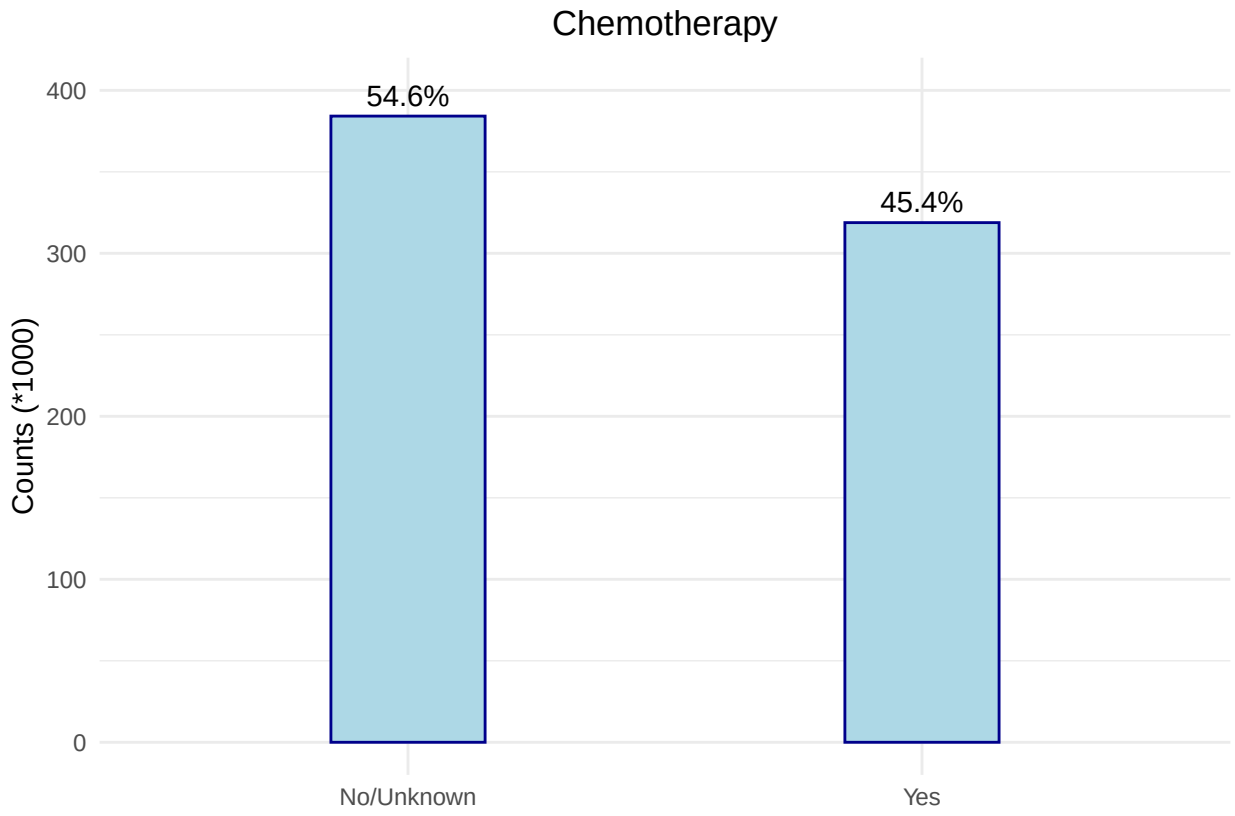


Chemotherapy (Binary)

```
# Create a summary table with counts and percentages
summary_table <- mydata_1 %>%
  group_by(Chemotherapy.recode..yes..no.unk.) %>%
  summarise(
    Count = n(),  # Count of each category
    Percentage = round((n() / nrow(mydata_1)) * 100, 1)  # Percentage of each category
  ) %>% ungroup()

# Display the summary table
kable(summary_table, caption = "")
```

Chemotherapy.recode..yes..no.unk.	Count	Percentage
No/Unknown	384179	54.6
Yes	318824	45.4

```r
# Calculate percentage frequencies
mydata_1 <- mydata_1 %>%
  group_by(Chemotherapy.recode..yes..no.unk.) %>%
  mutate(Count = n()) %>%
  mutate(Percentage = Count / nrow(mydata_1) * 100) %>% ungroup()

# Create barplot
ggplot(mydata_1, aes(x = Chemotherapy.recode..yes..no.unk.)) +
  geom_bar(aes(y = after_stat(count) / 1000), width = 0.3, fill = "lightblue", color = "darkblue") +
  geom_text(aes(y = after_stat(count) / 1000, label = sprintf("%.1f%%", Percentage)),
            stat = "count", vjust = -0.5, color = "black") +  # Add percentages on top of bars
  labs(title = "Chemotherapy", y = "Counts (*1000)", x = "") +
  scale_y_continuous(limits = c(0, 400)) +
  theme_minimal() +  # Optional: Use a minimal theme
  theme(plot.title = element_text(hjust = 0.5))  # Center the title
```

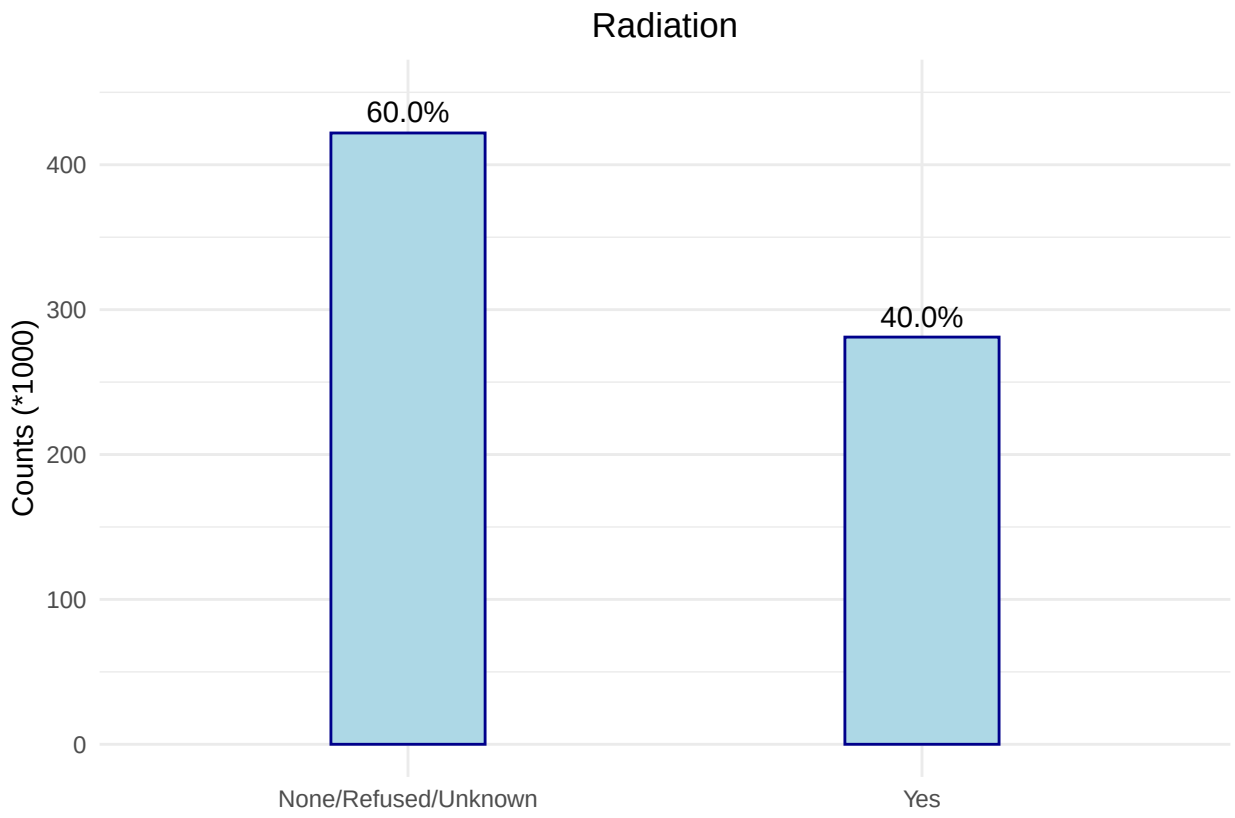


Radiation (Binary)

```r
# Create a summary table with counts and percentages
summary_table <- mydata_1 %>%
  group_by(radiation) %>%
  summarise(
    Count = n(),  # Count of each category
    Percentage = round((n() / nrow(mydata_1)) * 100, 1)  # Percentage of each category
  ) %>% ungroup()

# Display the summary table
kable(summary_table, caption = "")
```

radiation	Count	Percentage
None/Refused/Unknown	421944	60
Yes	281059	40

```r
# Calculate percentage frequencies
mydata_1 <- mydata_1 %>%
  group_by(radiation) %>%
  mutate(Count = n()) %>%
  mutate(Percentage = Count / nrow(mydata_1) * 100) %>% ungroup()

# Create barplot
ggplot(mydata_1, aes(x = radiation)) +
  geom_bar(aes(y = after_stat(count) / 1000), width = 0.3, fill = "lightblue", color = "darkblue") +
  geom_text(aes(y = after_stat(count) / 1000, label = sprintf("%.1f%%", Percentage)),
            stat = "count", vjust = -0.5, color = "black") +  # Add percentages on top of bars
  labs(title = "Radiation", y = "Counts (*1000)", x = "") +
  scale_y_continuous(limits = c(0, 450)) +
  theme_minimal() +  # Optional: Use a minimal theme
  theme(plot.title = element_text(hjust = 0.5))  # Center the title
```



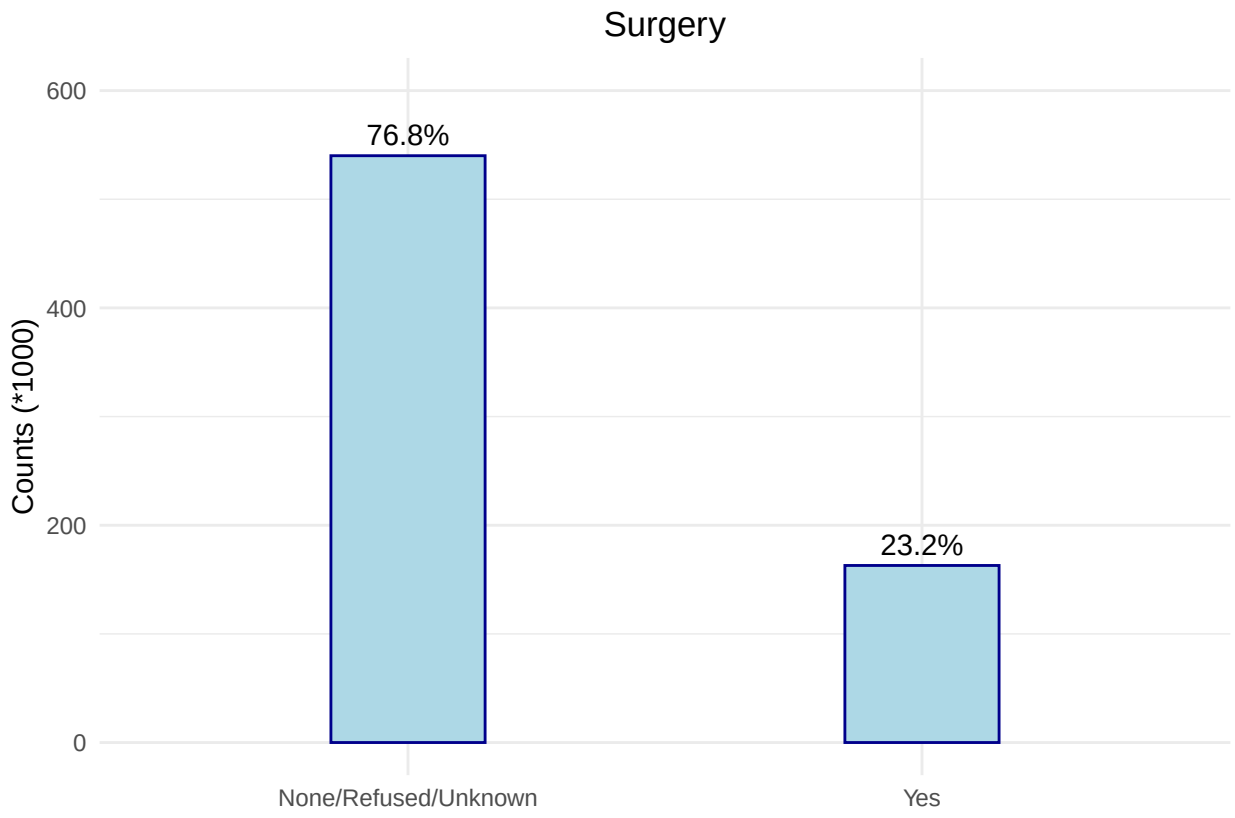
Surgery (Binary)

```
# Create a summary table with counts and percentages
summary_table <- mydata_1 %>%
  group_by(surgery) %>%
  summarise(
    Count = n(),  # Count of each category
    Percentage = round((n() / nrow(mydata_1)) * 100, 1)  # Percentage of each category
  ) %>% ungroup()

# Display the summary table
kable(summary_table, caption = "")
```

surgery	Count	Percentage
None/Refused/Unknown	540047	76.8
Yes	162956	23.2

```
# Calculate percentage frequencies
mydata_1 <- mydata_1 %>%
  group_by(surgery) %>%
  mutate(Count = n()) %>%
  mutate(Percentage = Count / nrow(mydata_1) * 100) %>% ungroup()
```

```r
# Create barplot
ggplot(mydata_1, aes(x = surgery)) +
  geom_bar(aes(y = after_stat(count) / 1000), width = 0.3, fill = "lightblue", color = "darkblue") +
  geom_text(aes(y = after_stat(count) / 1000, label = sprintf("%.1f%%", Percentage)),
            stat = "count", vjust = -0.5, color = "black") +  # Add percentages on top of bars
  labs(title = "Surgery", y = "Counts (*1000)", x = "") +
  scale_y_continuous(limits = c(0, 600)) +
  theme_minimal() +  # Optional: Use a minimal theme
  theme(plot.title = element_text(hjust = 0.5))  # Center the title
```



Vital status (Binary)

```r
# Create a summary table with counts and percentages
summary_table <- mydata_1 %>%
  group_by(Vital.status.recode..study.cutoff.used.) %>%
  summarise(
    Count = n(),  # Count of each category
    Percentage = round((n() / nrow(mydata_1)) * 100, 1)  # Percentage of each category
  ) %>% ungroup()

# Display the summary table
kable(summary_table, caption = "")
```

Vital.status.recode..study.cutoff.used.	Count	Percentage
Alive	115803	16.5
Dead	587200	83.5

Survival months (Continuous)

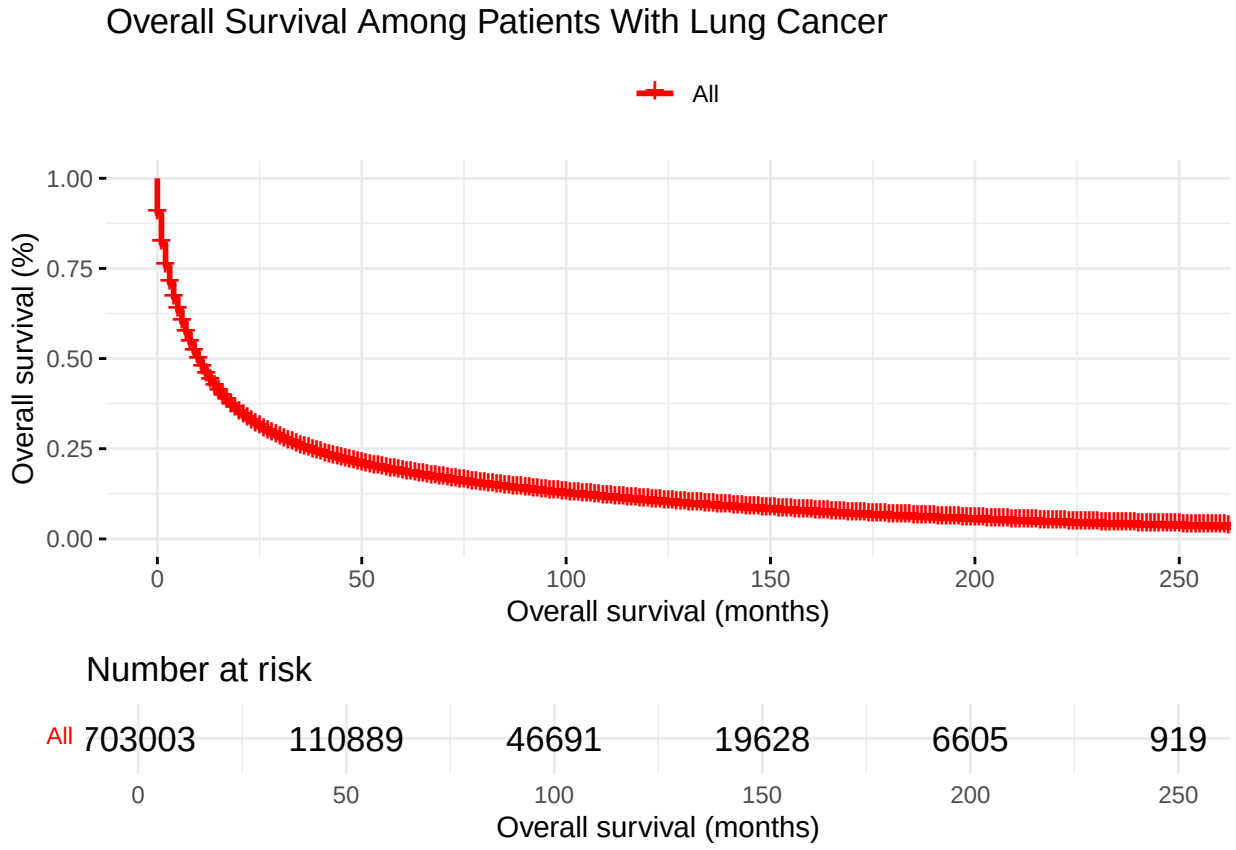
```
summary(mydata_1$Survival.months)  # Gives Min, 1st Qu., Median, Mean, 3rd Qu., Max
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    2.00    9.00   25.58   28.00  263.00
```

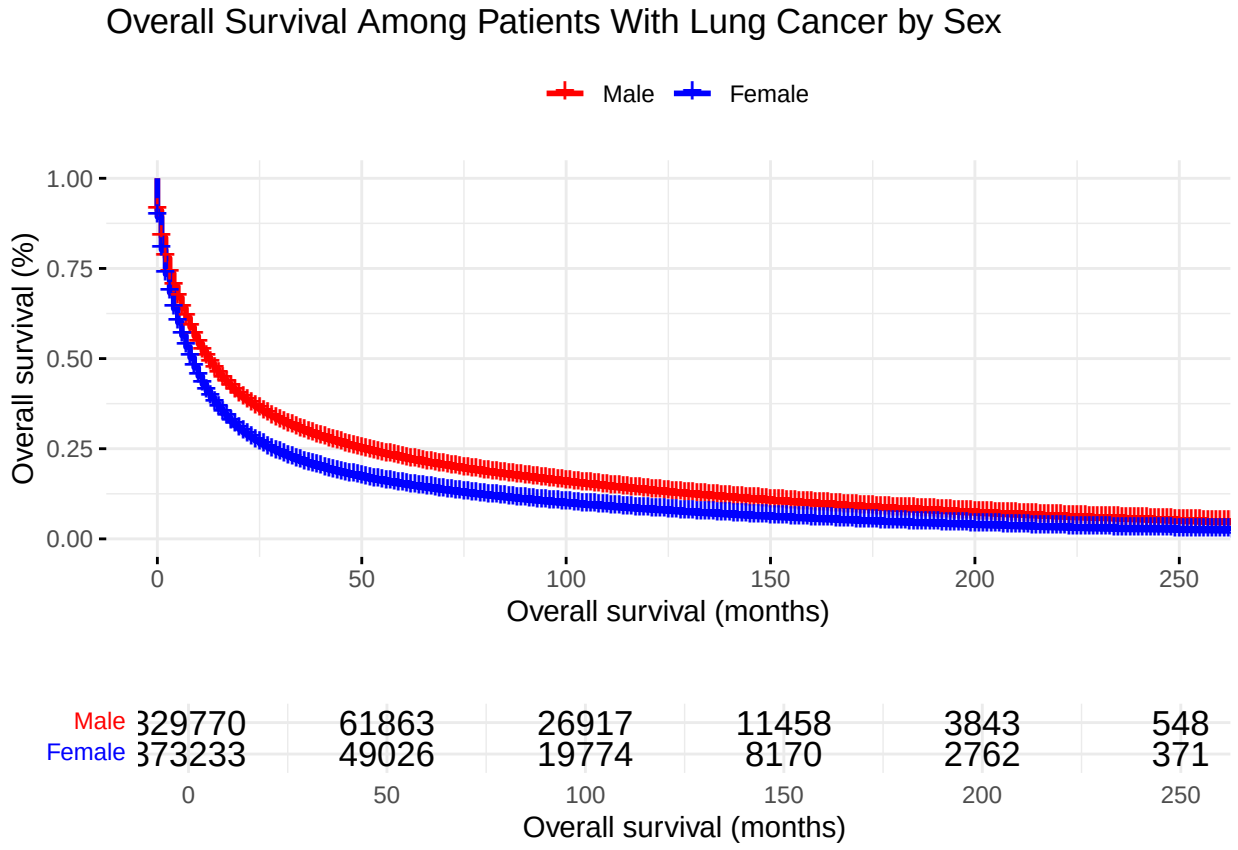
```
sd(mydata_1$Survival.months)       # Standard Deviation
```

```
## [1] 40.50034
```

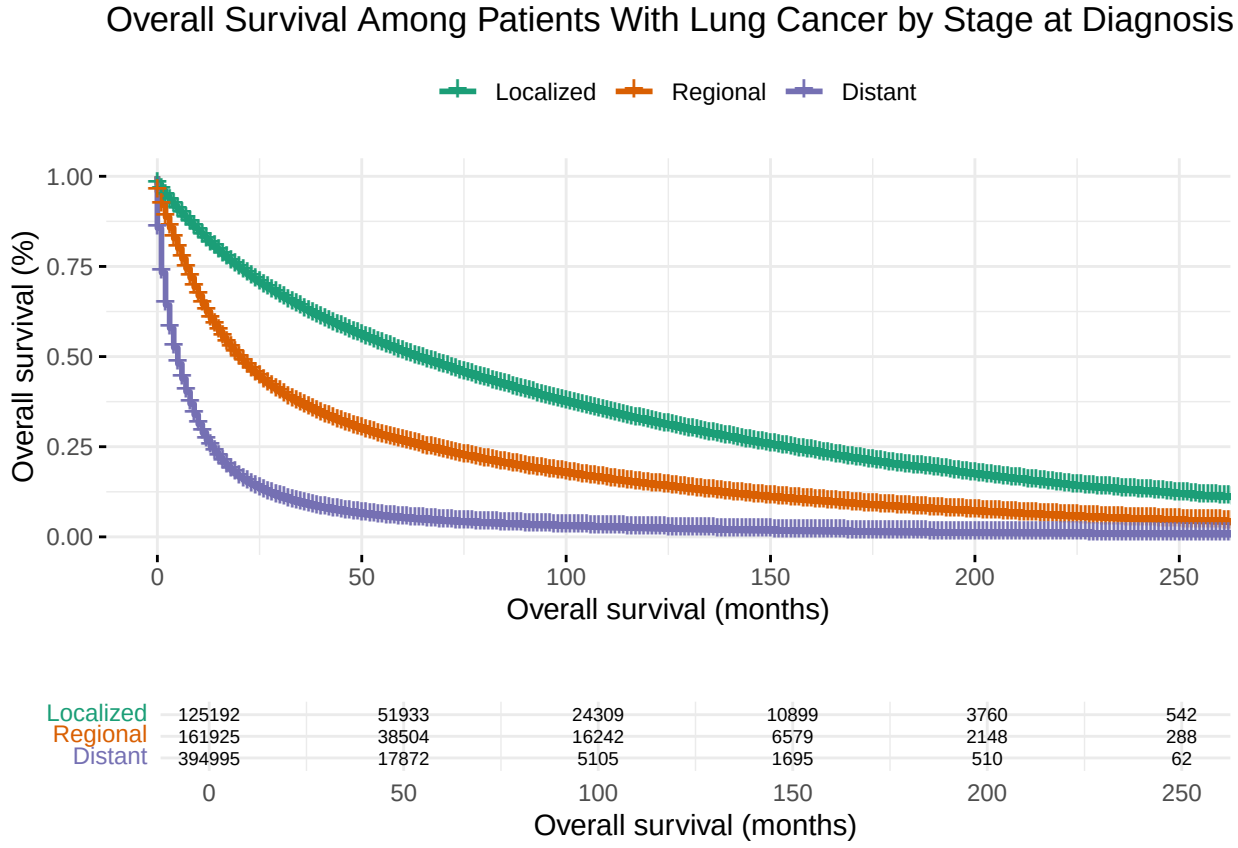
```
# Create a censoring variable: 1 if death occurred, 0 if censored (last follow-up)
mydata_1 <- mydata_1 %>%
  mutate(status = ifelse(Vital.status.recode..study.cutoff.used. == "Dead", 1, 0))  # 1 = death, 0 = ce

# Fit Kaplan-Meier survival model
km_fit <- survfit(Surv(Survival.months, status) ~ 1, data = mydata_1)

# Plot Kaplan-Meier survival curve by sex
ggsurvplot(km_fit,
           data = mydata_1,
           legend.title = "",
           xlab = "Overall survival (months)",
           ylab = "Overall survival (%)",
           title = "Overall Survival Among Patients With Lung Cancer",
           risk.table = TRUE,
           conf.int = FALSE,
           palette = c("red", "blue"),
           ggtheme = theme_minimal())
```



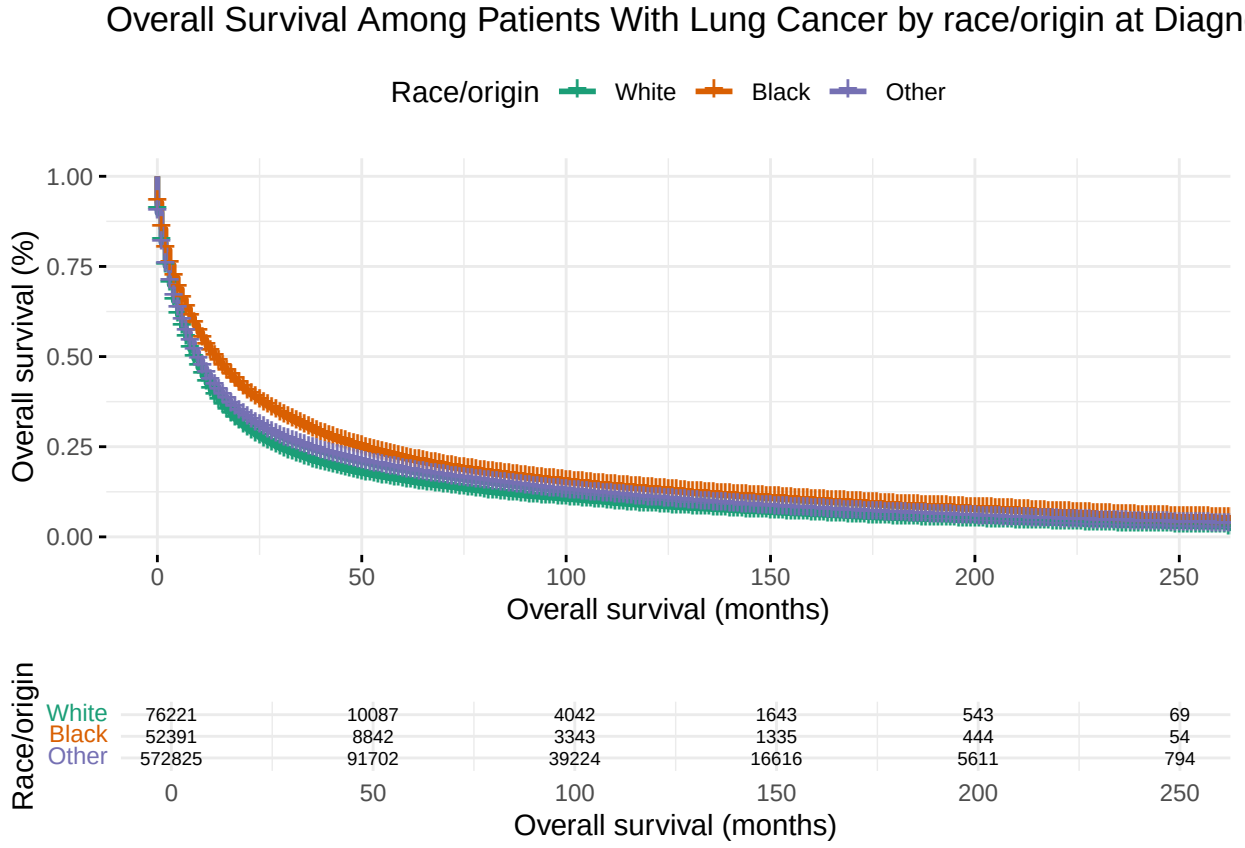

```r
# Fit Kaplan-Meier survival model, stratified by sex
km_fit_sex <- survfit(Surv(Survival.months, status) ~ Sex, data = mydata_1)

# Plot Kaplan-Meier survival curve by sex
ggsurvplot(km_fit_sex,
           data = mydata_1,
           legend.title = "",    # Legend title for groups
           legend.labs = c("Male", "Female"),  # Customize legend labels
           xlab = "Overall survival (months)",    # X-axis label
           ylab = "Overall survival (%)",  # Y-axis label
           title = "Overall Survival Among Patients With Lung Cancer by Sex",  # Title
           risk.table = TRUE,  # Show the risk table below the plot
           risk.table.title = "",
           conf.int = FALSE,
           palette = c("red", "blue"),  # Set colors for male and female
           ggtheme = theme_minimal())  # Set a minimal theme
```



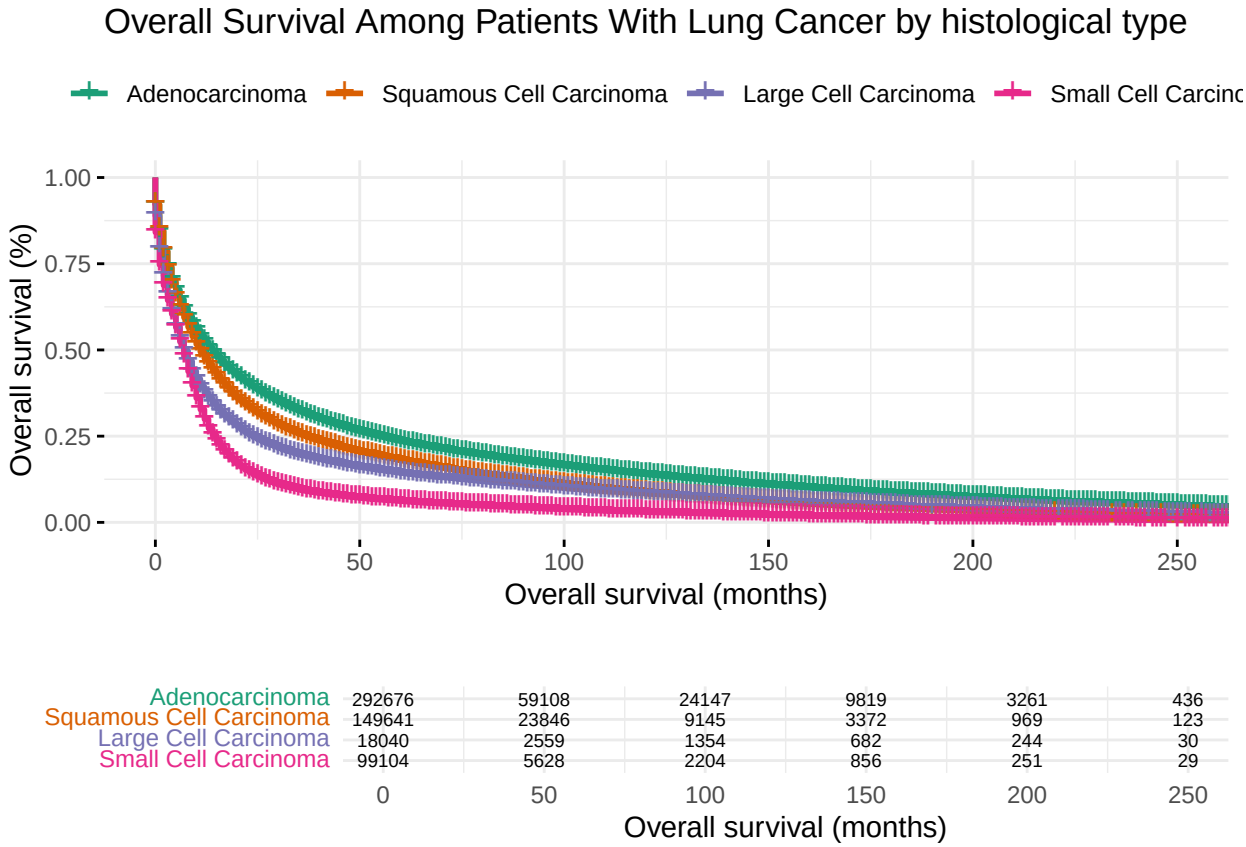

```r
# Subset the data to include only the desired stages
mydata_1_filtered <- mydata_1[mydata_1$combined_stage %in% c("Distant", "Localized", "Regional"), ]

# Set the levels of combined_stage in the desired order
mydata_1_filtered$combined_stage <- factor(mydata_1_filtered$combined_stage,
                                           levels = c("Localized", "Regional", "Distant"))

# Fit Kaplan-Meier survival model, stratified by combined stage
km_fit_stage <- survfit(Surv(Survival.months, status) ~ combined_stage, data = mydata_1_filtered)

# Plot Kaplan-Meier survival curve by stage
ggsurvplot(km_fit_stage,
           data = mydata_1_filtered,
           legend.title = "",
           legend.labs = c("Localized", "Regional", "Distant"),  # Legend labels in the desired order
           xlab = "Overall survival (months)",
           ylab = "Overall survival (%)",
           title = "Overall Survival Among Patients With Lung Cancer by Stage at Diagnosis",  # Adjust
           risk.table = TRUE,
           risk.table.title = "",
           conf.int = FALSE,
           palette = "Dark2",
           ggtheme = theme_minimal(),
           risk.table.fontsize = 2.5,
           risk.table.height = 0.25)
```




```r
# Subset the data to include only the desired race/origin
mydata_1_filtered <- mydata_1[mydata_1$Race.recode..White..Black..Other. %in% c("White", "Black", "Othe

# Set the levels of Race.recode..White..Black..Other. in the desired order
mydata_1_filtered$combined_stage <- factor(mydata_1_filtered$combined_stage,
                                           levels = c("White", "Black", "Other (American Indian/AK Nativ

# Fit Kaplan-Meier survival model, stratified by race/origin
km_fit_race <- survfit(Surv(Survival.months, status) ~ Race.recode..White..Black..Other., data = mydata_

# Plot Kaplan-Meier survival curve by race/origin
ggsurvplot(km_fit_race,
           data = mydata_1_filtered,
           legend.title = "Race/origin",
           legend.labs = c("White", "Black", "Other"),  # Legend labels in the desired order
           xlab = "Overall survival (months)",
           ylab = "Overall survival (%)",
           title = "Overall Survival Among Patients With Lung Cancer by race/origin at Diagnosis",  # A
           risk.table = TRUE,
           risk.table.title = "",
           conf.int = FALSE,
           palette = "Dark2",
           ggtheme = theme_minimal(),
           risk.table.fontsize = 2.5,
           risk.table.height = 0.25)
```




```r
# Subset the data to include only the desired histological type
mydata_1_filtered <- mydata_1[mydata_1$histological_type %in% c("Adenocarcinoma", "Squamous Cell Carcin

# Set the levels of histological_type in the desired order
mydata_1_filtered$histological_type <- factor(mydata_1_filtered$histological_type,
                                              levels = c("Adenocarcinoma", "Squamous Cell Carcinoma", "Large

# Fit Kaplan-Meier survival model, stratified by histological_type
km_fit_hs <- survfit(Surv(Survival.months, status) ~ histological_type, data = mydata_1_filtered)

# Plot Kaplan-Meier survival curve by histological_type
ggsurvplot(km_fit_hs,
           data = mydata_1_filtered,
           legend.title = "",
           legend.labs = c("Adenocarcinoma", "Squamous Cell Carcinoma", "Large Cell Carcinoma", "Small
           xlab = "Overall survival (months)",
           ylab = "Overall survival (%)",
           title = "Overall Survival Among Patients With Lung Cancer by histological type",  # Adjust t
           risk.table = TRUE,
           risk.table.title = "",
           conf.int = FALSE,
           palette = "Dark2",
           ggtheme = theme_minimal(),
           risk.table.fontsize = 2.5,
           risk.table.height = 0.27)
```



Cause of death 1

```
# Create a summary table with counts and percentages
summary_table <- mydata_1 %>%
  group_by(SEER.other.cause.of.death.classification) %>%
  summarise(
    Count = n(),  # Count of each category
    Percentage = round((n() / nrow(mydata_1)) * 100, 1)  # Percentage of each category
  ) %>% ungroup()

# Display the summary table
kable(summary_table, caption = "")
```

SEER.other.cause.of.death.classification	Count	Percentage
Alive or dead due to cancer	615314	87.5
Dead (attributable to causes other than this cancer dx)	81854	11.6
Dead (missing/unknown COD)	5835	0.8

Cause of death 2

```r
# Create a summary table with counts and percentages
summary_table <- mydata_1 %>%
  group_by(SEER.cause.specific.death.classification) %>%
  summarise(
    Count = n(),  # Count of each category
    Percentage = round((n() / nrow(mydata_1)) * 100, 1)  # Percentage of each category
  ) %>% ungroup()

# Display the summary table
kable(summary_table, caption = "")
```

SEER.cause.specific.death.classification	Count	Percentage
Alive or dead of other cause	197657	28.1
Dead (attributable to this cancer dx)	499511	71.1
Dead (missing/unknown COD)	5835	0.8

Cause of death 3

```r
# Create a summary table with counts and percentages
summary_table <- mydata_1 %>%
  group_by(COD.to.site.recode.ICD.O.3.2023.Revision.Expanded..1999..) %>%
  summarise(
    Count = n(),  # Count of each category
    Percentage = round((n() / nrow(mydata_1)) * 100, 1)  # Percentage of each category
  ) %>% ungroup() %>% arrange(desc(Count))

# Display the summary table
kable(summary_table, caption = "")
```

COD.to.site.recode.ICD.O.3.2023.Revision.Expanded..1999..	Count	Percentage
Lung And Bronchus	473467	67.3
Alive	115803	16.5
Miscellaneous Neoplasms	16893	2.4
Chronic Obstructive Pulmonary Disease and Allied Conditions	15838	2.3
Ischemic heart disease	14416	2.1
Other COD	12495	1.8
Other and unspecified disorders of the circulatory system	6801	1.0
State DC not available or state DC available but no COD	5835	0.8
Cerebrovascular diseases	4318	0.6
Pneumonia and Influenza	2751	0.4
Accidents and Adverse Effects	2344	0.3
Septicemia	2101	0.3
Hypertensive disease	1866	0.3
COVID (2020+ only)	1602	0.2
Diabetes Mellitus	1590	0.2
Symptoms, Signs and Ill-Defined Conditions	1531	0.2
Other infectious and Parasitic Diseases incl HIV	1466	0.2
Alzheimers (ICD-9 and ICD-10 only)	1346	0.2

COD.to.site.recode.ICD.O.3.2023.Revision.Expanded..1999..	Count	Percentage
Pancreas	1317	0.2
Colon And Rectum (Excluding Appendix)	1278	0.2
Nephritis, Nephrotic Syndrome and Nephrosis	1265	0.2
Breast	1186	0.2
Diseases of arteries, arterioles and capillaries	1101	0.2
Liver	1058	0.2
Brain (Malignant)	1006	0.1
Benign and Borderline: All Other sites	889	0.1
Suicide and Self-Inflicted Injury	795	0.1
Esophagus	712	0.1
Pulmonary heart disease and diseases of pulmonary circulation	695	0.1
Miscellaneous Hematopoietic Neoplasms	675	0.1
Chronic Liver Disease and Cirrhosis	586	0.1
Prostate	527	0.1
Kidney Parenchyma	474	0.1
Urinary Bladder	471	0.1
Other B-cell leukemia/lymphomas or Lymphoma, NOS	460	0.1
Soft Tissue	421	0.1
Stomach	357	0.1
Acute Myeloid Leukemias	345	0.0
Brain, CNS Other and Intracranial Gland (Benign and Borderline)	302	0.0
Mesothelioma	256	0.0
Other Non-Epithelial Skin	250	0.0
Ovary	236	0.0
Stomach and Duodenal Ulcers	230	0.0
Complications of medical and surgical care (Y40-Y84, Y88) (ICD-10 only, 1999+)	226	0.0
Heart, Mediastinum And Pleura	219	0.0
Larynx	215	0.0
Plasma Cell Neoplasms	198	0.0
Other Leukemias	180	0.0
Myelodysplastic Syndromes	177	0.0
Melanoma Of The Skin	174	0.0
Bones And Joints	173	0.0
Trachea, And Respiratory Other	137	0.0
Intrahepatic Bile Duct	134	0.0
Adrenal Gland	121	0.0
Pharynx And Oral Cavity Other	121	0.0
Corpus	113	0.0
Digestive Other	97	0.0
Thyroid	90	0.0
Chronic lymphocytic leukemia (CLL)/Small lymphocytic lymphoma	86	0.0
Congenital Anomalies	73	0.0
Small Intestine	73	0.0
Oropharynx	68	0.0
Tongue Anterior	66	0.0
Myeloproliferative/Myelodysplastic syndromes, including MDS/MPN overlap	59	0.0
Homicide and Legal Intervention	52	0.0
Large B-cell lymphoma	52	0.0
Complications of Pregnancy, Childbrith, Puerperium	45	0.0
Cervix	44	0.0
Mouth Other	41	0.0
Gallbladder	40	0.0

COD.to.site.recode.ICD.O.3.2023.Revision.Expanded..1999..	Count	Percentage
In situ neoplasms	39	0.0
Thymus	38	0.0
Meninges (Malignant)	37	0.0
Endocrine Other	34	0.0
Retroperitoneum And Peritoneum	32	0.0
Biliary Other	31	0.0
Nasopharynx	30	0.0
Major Salivary Glands	29	0.0
Precursor Lymphoid Neoplasms	28	0.0
CNS Other (Malignant)	26	0.0
Anus, Anal Canal And Anorectum	22	0.0
Hodgkin Lymphomas	21	0.0
Other T and NK-cell leukemias/lymphomas	21	0.0
Extrahepatic Bile Ducts	15	0.0
Testis	14	0.0
Ureter	13	0.0
Hypopharynx	11	0.0
Kidney Renal Pelvis	11	0.0
Sinus Other	11	0.0
Urinary Other	11	0.0
Vulva	10	0.0
Appendix	9	0.0
Kaposi Sarcoma	8	0.0
Ampulla Of Vater	7	0.0
Certain Conditions Originating in Perinatal Period	7	0.0
Nasal Cavity And Paranasal Sinuses	7	0.0
Urethra	6	0.0
Fallopian Tube	5	0.0
Gum	5	0.0
Lip	5	0.0
Penis	5	0.0
Vagina	5	0.0
Adnexa Other And Genital Female Other	4	0.0
Intracranial Gland (Malignant)	4	0.0
Eye And Orbit	3	0.0
Floor Of Mouth	3	0.0
Placenta	3	0.0
Buccal Mucosa	1	0.0
Genital Male Other	1	0.0
Mycosis Fungoides/Sezary Syndrome	1	0.0
Parathyroid	1	0.0