Brain Cancer Survival:

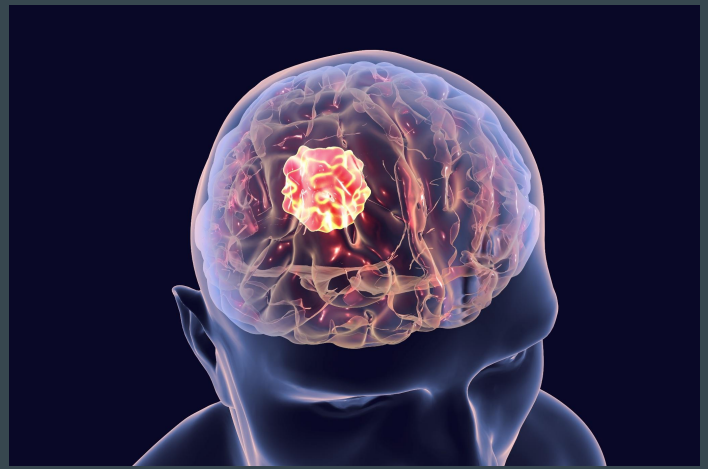
Understanding Factors Contributing to Survival of Patients Using SEER Data

Daniel Cerkoney and Anthony Young

Erdos Institute Data Science Boot Camp May 2024

Introduction/Motivation

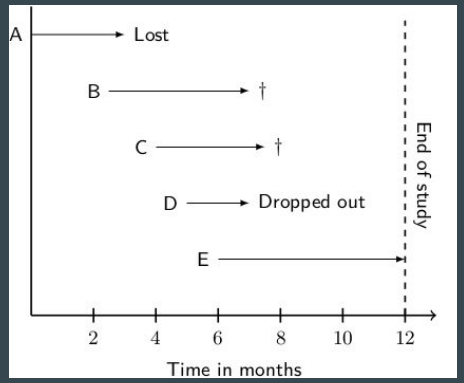
- Cancer is the #2 leading cause of death in the US.
- Cancer treatment is complex and highly individualized.
- Brain cancer is one of the hardest varieties to treat.
- Surveillance, Epidemiology, and End Results (SEER) Program
    - Large dataset of anonymized cancer listings
    - Compiled from several cancer registries
    - Covers 42% of US population
    - 1975-present
    - Tracks a large number of covariates including:
        - Demographics
        - Tumor characteristics
        - Diagnostic results
        - And much more!

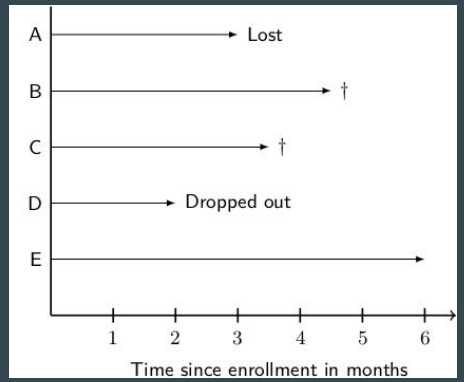


Survival Analysis

- Goal: predict survival S(t) as a function of time
  - Key distinctions from standard multivariate regression:
    - Censoring: lost to follow-up or end of study
    - Censoring ⇒ time dependence
- Methods:
  - Kaplan-Meier estimator
    - Survival estimates from raw data (no fitting required)
  - Multivariate Cox regression
    - Survival estimates account for covariates
    - Central assumption: survival functions are proportional throughout time

$$S(t\,|X) = S_0(t)\exp\left(X^T\beta\right)$$

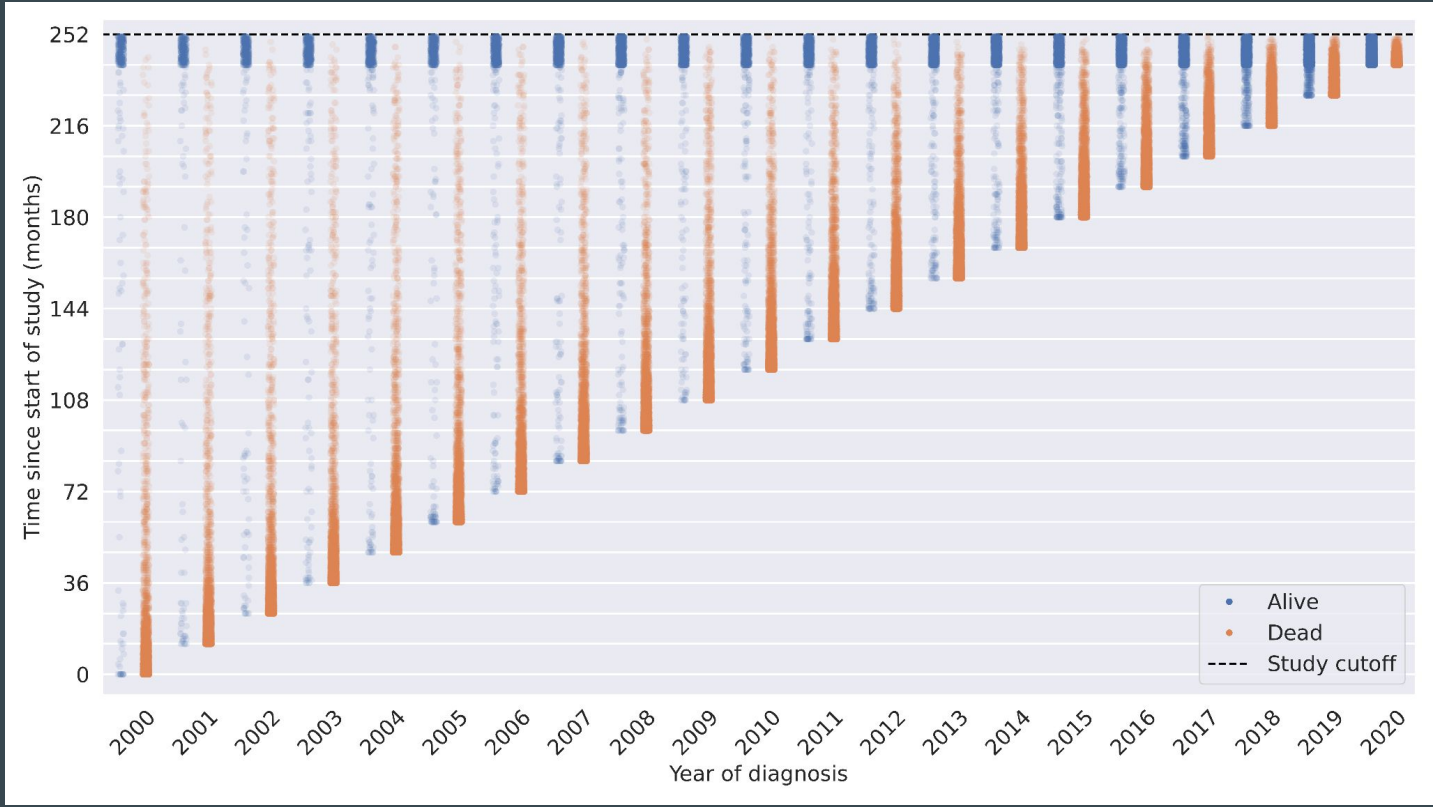




Data Collection

- Analyzed the SEER 17 database (Nov. 2022 submission)
  - Covers ~26.5% of the US population based on a 2020 population census
  - Includes cases documented between 2000 and 2020
  - 9,208,295 total tumor records
- Used the SEER*Stat software package for data collection
- Extracted case info and survival data for all 76,327 brain tumor entries
- Restricted the study cohort to patients with a single tumor ⇒ 74,332 cases

Data Collection



Data Preprocessing & Feature Engineering

- Redundant features in SEER database
- Selected a subset of all features to work with
- Fill missing values & add features to track them
- Most categories are categorical (i.e., demographics, diagnostic encodings)
  - One-hot encoding creates many features.
  - We drop categories that are too rare.
- Automatically drop features with high correlations to other features
- 111 features after preprocessing

Model Training

- Used the lifelines survival analysis library
- Model scoring via partial log-likelihood
  - Depends on the censoring pattern
  - Generalizes maximum likelihood estimation to survival analysis
- Employ L1 and L2 regularization to improve training performance & stability
- 3-fold cross-validation of regularization hyperparameters
- Stratification using age standards for survival analysis
- Assess goodness-of-fit via the concordance index (c-index)
- Model performance: c-index = 0.726

Results

Covariate	Hazard Ratio	Lower 95% CI	Upper 95% CI	p-value
Glioblastoma	1.636	1.580	1.694	<0.0005
"Months from diagnosis to treatment" data missing	1.252	1.236	1.268	<0.0005
ICD-O-3 Code: 9421	0.864	0.830	0.899	<0.0005
ICD-O-3 Code: 9064	0.897	0.872	0.923	<0.0005
Localized tumor	0.906	0.889	0.923	<0.0005
Year of diagnosis	0.917	0.890	0.945	<0.0005
Oligodendroglioma, IDH-mutant & 1 p/19q co-deleted	0.920	0.880	0.963	<0.0005
Primary Site: Cerebrum	1.084	1.073	1.095	<0.0005
Diffuse astrocytoma and anaplastic astrocytoma	1.080	1.049	1.111	<0.0005
ICD-O-3 Code: 8000	1.079	1.062	1.095	<0.0005
ICD-O-3 Code: 9450	0.930	0.909	0.951	<0.0005
Diffuse astrocytoma, IDH-mutant	0.930	0.903	0.957	<0.0005
Other astrocytic tumor	0.930	0.895	0.966	<0.0005
Ependymal tumor	0.932	0.861	1.009	0.084
Summary Stage: Regional/Not otherwise specified	1.073	1.058	1.088	<0.0005
ICD-O-3 Code: 9401	1.070	1.053	1.087	<0.0005
Tumor Grade I (Well differentiated)	0.937	0.925	0.949	<0.0005
Tumor Grade II (Moderately differentiated)	0.938	0.926	0.949	<0.0005
ICD-O-3 Code: 9391	0.938	0.873	1.008	0.082
Sex: Female	0.939	0.930	0.948	<0.0005

Results

Covariate	Hazard Ratio	Lower 95% CI	Upper 95% CI	p-value
Glioblastoma	1.636	1.580	1.694	<0.0005
"Months from diagnosis to treatment" data missing	1.252	1.236	1.268	<0.0005
ICD-O-3 Code: 9421	0.864	0.830	0.899	<0.0005
ICD-O-3 Code: 9064	0.897	0.872	0.923	<0.0005
Localized tumor	0.906	0.889	0.923	<0.0005
Year of diagnosis	0.917	0.890	0.945	<0.0005
Oligodendroglioma, IDH-mutant & 1 p/19q co-deleted	0.920	0.880	0.963	<0.0005
Primary Site: Cerebrum	1.084	1.073	1.095	<0.0005
Diffuse astrocytoma and anaplastic astrocytoma	1.080	1.049	1.111	<0.0005
ICD-O-3 Code: 8000	1.079	1.062	1.095	<0.0005
ICD-O-3 Code: 9450	0.930	0.909	0.951	<0.0005
Diffuse astrocytoma, IDH-mutant	0.930	0.903	0.957	<0.0005
Other astrocytic tumor	0.930	0.895	0.966	<0.0005
Ependymal tumor	0.932	0.861	1.009	0.084
Summary Stage: Regional/Not otherwise specified	1.073	1.058	1.088	<0.0005
ICD-O-3 Code: 9401	1.070	1.053	1.087	<0.0005
Tumor Grade I (Well differentiated)	0.937	0.925	0.949	<0.0005
Tumor Grade II (Moderately differentiated)	0.938	0.926	0.949	<0.0005
ICD-O-3 Code: 9391	0.938	0.873	1.008	0.082
Sex: Female	0.939	0.930	0.948	<0.0005

Prediction: a glioblastoma diagnosis is associated with a 64% higher risk of death

Results

Covariate	Hazard Ratio	Lower 95% CI	Upper 95% CI	p-value
Glioblastoma	1.636	1.580	1.694	<0.0005
"Months from diagnosis to treatment" data missing	1.252	1.236	1.268	<0.0005
ICD-O-3 Code: 9421	0.864	0.830	0.899	<0.0005
ICD-O-3 Code: 9064	0.897	0.872	0.923	<0.0005
Localized tumor	0.906	0.889	0.923	<0.0005
Year of diagnosis	0.917	0.890	0.945	<0.0005
Oligodendroglioma, IDH-mutant & 1 p/19q co-deleted	0.920	0.880	0.963	<0.0005
Primary Site: Cerebrum	1.084	1.073	1.095	<0.0005
Diffuse astrocytoma and anaplastic astrocytoma	1.080	1.049	1.111	<0.0005
ICD-O-3 Code: 8000	1.079	1.062	1.095	<0.0005
ICD-O-3 Code: 9450	0.930	0.909	0.951	<0.0005
Diffuse astrocytoma, IDH-mutant	0.930	0.903	0.957	<0.0005
Other astrocytic tumor	0.930	0.895	0.966	<0.0005
Ependymal tumor	0.932	0.861	1.009	0.084
Summary Stage: Regional/Not otherwise specified	1.073	1.058	1.088	<0.0005
ICD-O-3 Code: 9401	1.070	1.053	1.087	<0.0005
Tumor Grade I (Well differentiated)	0.937	0.925	0.949	<0.0005
Tumor Grade II (Moderately differentiated)	0.938	0.926	0.949	<0.0005
ICD-O-3 Code: 9391	0.938	0.873	1.008	0.082
Sex: Female	0.939	0.930	0.948	<0.0005

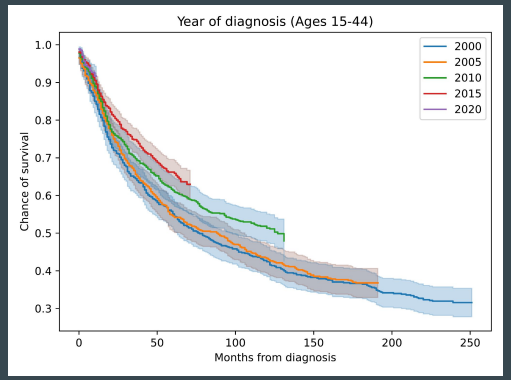
Prediction: 1p/19q co-deletion is associated with an 8% lower risk of death

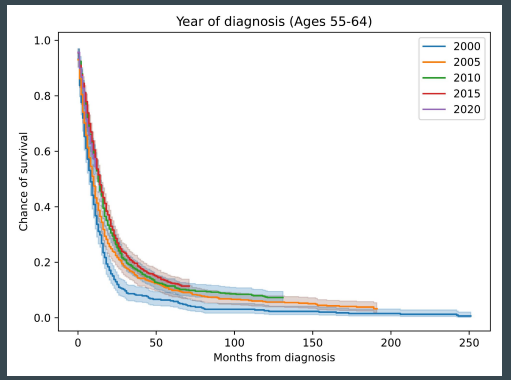
Survival adjusted for year of diagnosis

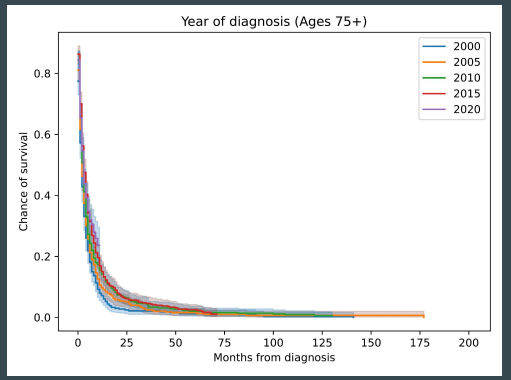
Kaplan-Meier:

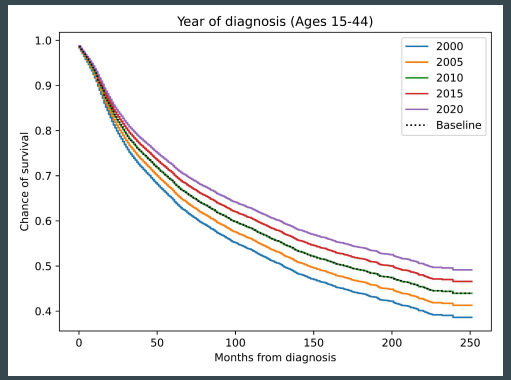
Adjust for covariates

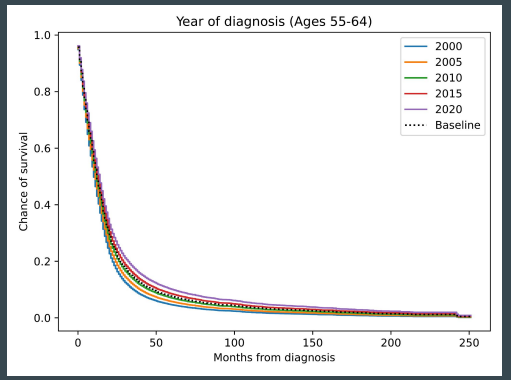
Cox regression:

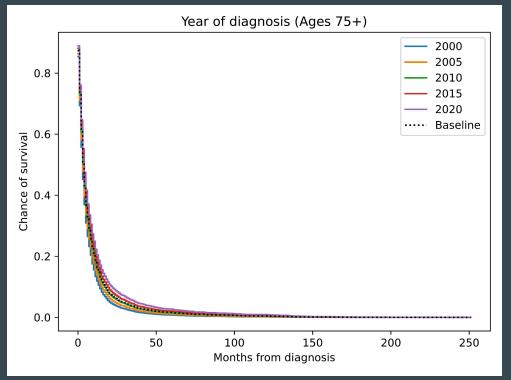










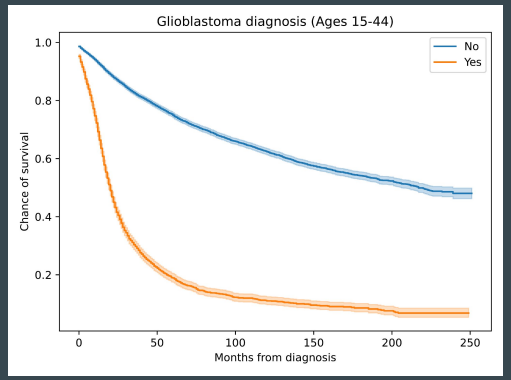


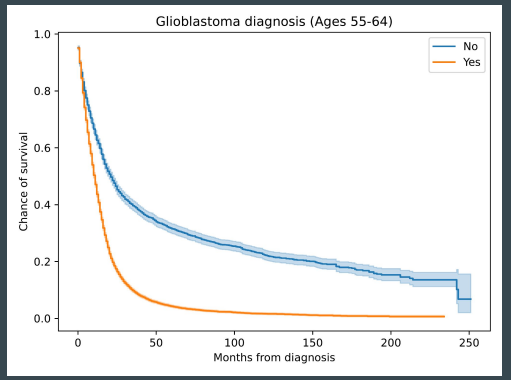
Survival adjusted for glioblastoma diagnosis

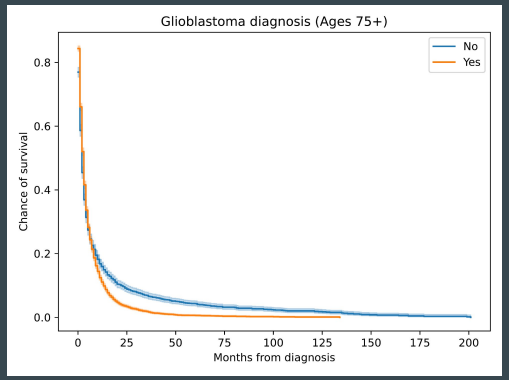
Kaplan-Meier:

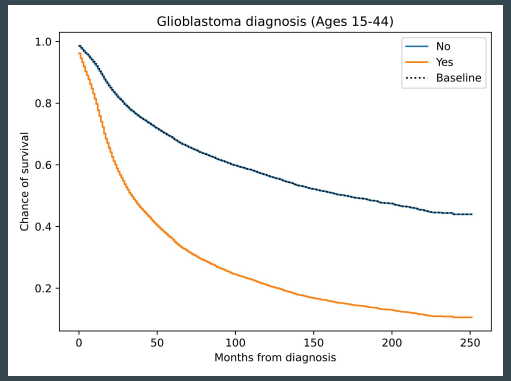
Adjust for covariates

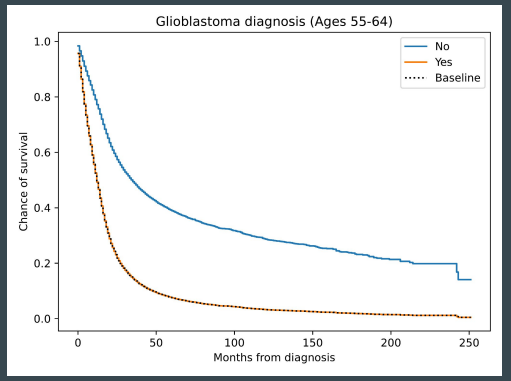
Cox regression:

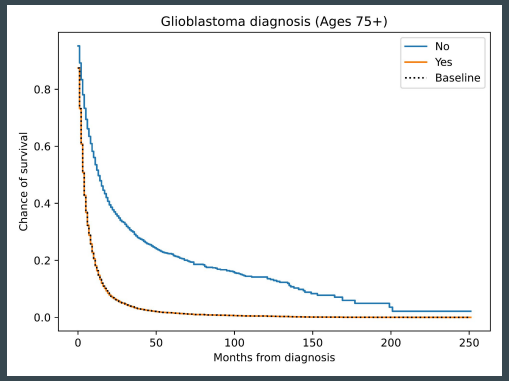












Summary & Future Work

- Analyzed brain tumor case data from the SEER cancer registry
- Predicted survival using multivariate Cox regression
- Future work:
  - Quantify proportional hazards assumption validity
  - Investigate factors that improve glioblastoma survival
  - Comparison with ensemble and deep learning methods:
    - Random survival forests
    - Survival support vector machines
    - Deep learning methods for survival analysis

Overview

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

Project objective:

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do

Understanding the market

Target audience

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt

The competition:

- Lorem ipsum
- Dolor sit amet

Market trends

Trend 1

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor

Client Implications:

- Incididunt ut labore et dolore
- Consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore

Trend 2

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor

Client Implications:

- Incididunt ut labore et dolore
- Consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore

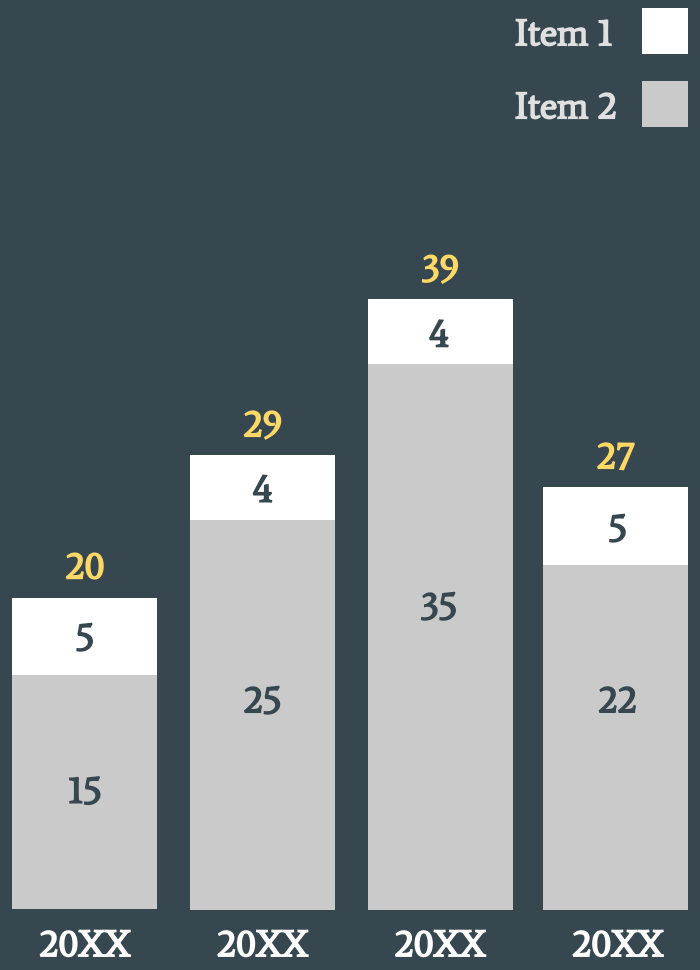
Trend analysis

Findings

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor

Client Implications:

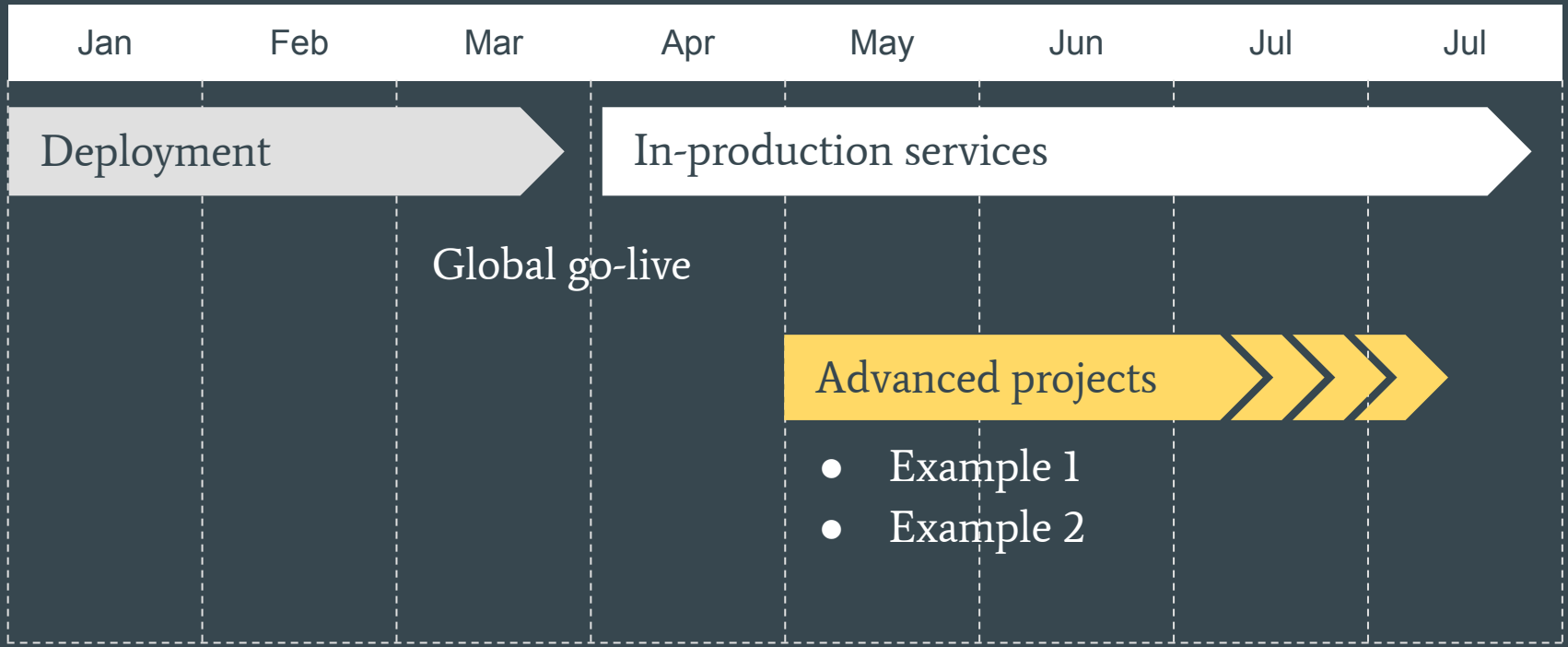
- Incididunt ut labore et dolore
- Consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore



Proposed deliverables

Deliverable 1	• Lorem ipsum dolor sit amet • Sed do eiusmod tempor incididunt ut labore
Deliverable 2	• Lorem ipsum dolor sit amet • Sed do eiusmod tempor incididunt ut labore
Deliverable 3	• Lorem ipsum dolor sit amet • Sed do eiusmod tempor incididunt ut labore
Deliverable 4	• Lorem ipsum dolor sit amet • Sed do eiusmod tempor incididunt ut labore

Timeline


Jan
Feb
Mar
Apr
May
Jun
Jul
Jul
Deployment
In-production services
Global go-live
Advanced projects
Example 1
Example 2

The Team

Wendy Writer, CEO

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor

Ronny Reader, CFO

Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat

Abby Author, CTO

Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur

Berry Books, CPO

Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum