

BUILDING HIGH-QUALITY SDTM DATASETS FROM REAL-WORLD DATA: LESSONS AND SOLUTIONS

Peishan(Jenny) Han, Lorraine Fang, Lixiao Chen, Kelvin Lin, Lu Gao

Bristol Myers Squibb, Lawrence Township, NJ, USA

March, 2026

BACKGROUND OF RWE AND RWD

- **Real-World Data (RWD):**

Data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources.

- **Real-World Evidence (RWE):**

The clinical evidence about the usage and potential benefits or risks of a medical product derived from analysis of RWD.

- Evidence from traditional clinical trials will not be considered RWE. However, various hybrid or pragmatic trial designs and observational studies could generate RWE.
- **Why RWE?**

Improve efficiencies of drug development program.

"To work with RWD across multiple sources, data may need to be put into a common format, sometimes referred to as a common data model(CDM), with common representation (terminologies, vocabularies, coding schemes). FDA recognizes the importance of developing data standards to maximize the utility of RWD and is working on identifying relevant standards and methodologies for collection and analysis of RWD. FDA has already been active in developing data standards for regulatory use and will continue to expand its work in this area. FDA will consider data standards along with the other critical aspects of the RWE Program. "

REGULATORY EXPECTATIONS FOR SUBMISSIONS USING REAL-WORLD DATA

Transparency and Data Provenance

Regulators require clear articulation of data element provenance, including definitions and transformation logic for RWD sources.

Comprehensive Metadata Documentation

Define-XML and Study Data Reviewer's Guide detail derivations, challenges, and interpretive steps in RWD standardization.

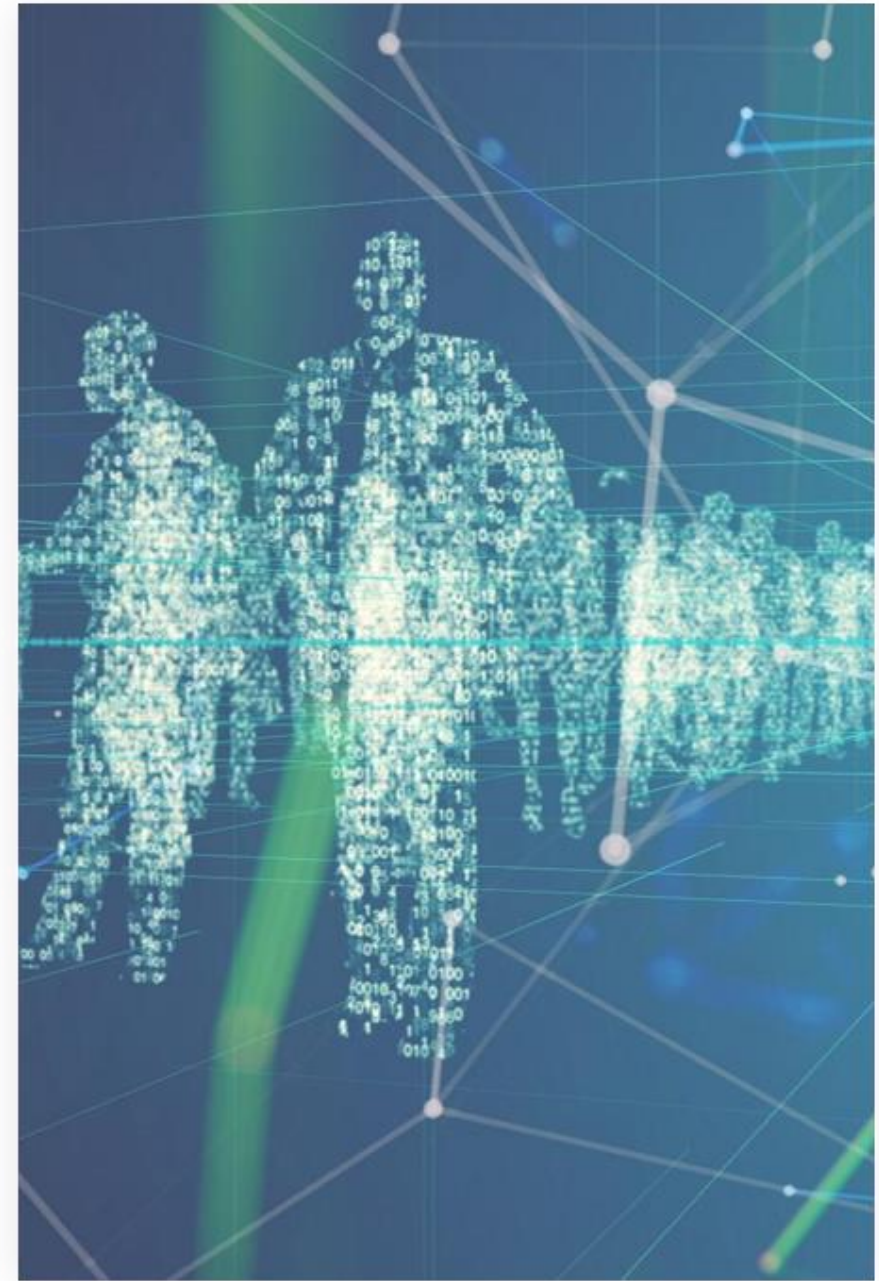
Justification of Data Sources

Sponsors must justify selection of data sources, curation steps, and harmonization rules across varied datasets.

Validation and Cross-functional Collaboration

Early and repeated validation with tools like Pinnacle 21 is essential, supported by teamwork across disciplines.

KEY CHALLENGES OF STANDARDIZING REAL-WORLD DATA



Data Quality Assessment

RWD often contains a high degree of missingness or lacks key assessments relevant to inclusion/exclusion criteria, baseline characteristics, prognostic factors, and longitudinal outcomes.

Large Databases and Subject Identification

ECA studies typically begin with very large RWD databases (e.g., >10,000 subjects) when raw data are not pre-curated for the study. Determining which patients should enter the SDTM pipeline requires early application of high-level inclusion/exclusion criteria.

Extensive and Heterogeneous Lab Data

Sponsors must justify selection of data sources, curation steps, and harmonization rules across varied datasets.

Free-Text Baseline Assessments

Baseline characteristics such as disease stage, plasmacytomas, CNS involvement, and treatment history are often captured as free text, with inconsistent variable naming across data sources.

Comorbidities Across Multiple Data files

Comorbidities may be recorded across multiple datasets within the same data source, including medical history, adverse events, diagnoses, and hospitalization records.

Inconsistent Missing Date Rules and Abnormalities in Treatment Timing

Different data vendors apply different date imputation rules. RWD also frequently contains implausible dates (e.g., treatment start dates occurring after treatment end dates).

Line of Therapy (LoT) Disagreement Across Stakeholders

Clinical teams often disagree with vendor-derived LoT definitions due to:

1. Variations in regimen construction
2. Inconsistent handling of treatment gaps or overlaps
3. Complex post-transplant treatment periods

Outcomes Distributed Across Many Sources

Outcome data may be distributed across multiple sources, including response assessments, reasons for regimen change, treatment discontinuation records, and clinical outcome files.

Domain Derivation Order Is Not the Same as in Clinical Trials

In clinical trials, SDTM programming typically follows a standard sequence in which inclusion/exclusion and demographics are collected on eCRFs at screening, allowing IE and DM domains to be derived early. In contrast, the complex I/E definitions in RWE studies and the need to determine an index date require adjustments to the order of SDTM domain derivation.

SOLUTION PRINCIPLES FOR HIGH-QUALITY SDTM FROM RWD



Early Alignment on High-Level I/E Criteria

- Engage statisticians and clinicians early to define broad inclusion/exclusion criteria that can be applied prior to SDTM mapping, based on data availability and study design. This reduces the subject pool to a manageable size while preserving key study requirements. Data quality should then be assessed on this subset to support customized data extraction requests from vendors.

Variable	Priority
CNS Involvement for all 172 patients	Can be removed
Plasmacytomas / extramedullary disease for all 172 patients any time prior to the index date	Must have
Bone Lesions - for 60 patients that did not have bone lytic lesions present in Q42023 (last time this was abstracted) any time prior to the index date	Must have
ECOG up to 3 months prior to index date - for 60 patients to supplement structured data	Must have, please expand to 3-6 months window (fees will be adjusted)
Comorbidities - congestive heart failure (CHF), myocardial infarction (MI), angina, pericardial, ischemic attack, ventricular arrhythmia, atrial arrhythmia, QT interval prolongation within 6 months prior to the index date	Must have
Comorbidities: LVEF for all 172 patients within 6 months prior to the index date	Must have
Comorbidities - pregnancy and breastfeeding	Can be removed
Comorbidities - history of HIV	Can be removed
B2M (as proxy for ISS stage) for 126 patients within 6 months prior to the index date	Must have

Identification of Key Lab Tests

– Among hundreds of lab tests, focus on laboratory tests relevant to the indication, inclusion/exclusion criteria, and prognostic factors. Create a harmonization Excel file documenting:

- Free-text variants
- Standardized LBTEST/LBTESTCD values
- Unit conversion rules

PARAM	LBTESTCD	LBTEST	LBSPEC	LBCAT	UNIT	SI_UNIT	SI_CONV	SI_ROUND
WBC	WBC	Leukocytes	BLOOD	HEMATOLOGY II	10 ³ /uL	10 ⁹ /L	1	0.01
WBC	WBC	Leukocytes	BLOOD	HEMATOLOGY II	K/UL	10 ⁹ /L	1	0.01
WBC	WBC	Leukocytes	BLOOD	HEMATOLOGY II	THOU/CU MM	10 ⁹ /L	1	0.01
WBC	WBC	Leukocytes	BLOOD	HEMATOLOGY II	THOU/UL	10 ⁹ /L	1	0.01
WBC	WBC	Leukocytes	BLOOD	HEMATOLOGY II	THOU/uL	10 ⁹ /L	1	0.01
WBC	WBC	Leukocytes	BLOOD	HEMATOLOGY II	k/uL	10 ⁹ /L	1	0.01
WBC	WBC	Leukocytes	BLOOD	HEMATOLOGY II	k/ul	10 ⁹ /L	1	0.01
WBC	WBC	Leukocytes	BLOOD	HEMATOLOGY II	k?ul	10 ⁹ /L	1	0.01
WBC	WBC	Leukocytes	BLOOD	HEMATOLOGY II	x10**3/uL	10 ⁹ /L	1	0.01
WBC	WBC	Leukocytes	BLOOD	HEMATOLOGY II	x10^3	10 ⁹ /L	1	0.01
WBC	WBC	Leukocytes	BLOOD	HEMATOLOGY II	x10^3/uL	10 ⁹ /L	1	0.01
WBC	WBC	Leukocytes	BLOOD	HEMATOLOGY II	x10^3/ul	10 ⁹ /L	1	0.01
WBC	WBC	Leukocytes	BLOOD	HEMATOLOGY II	x10^9/L	10 ⁹ /L	1	0.01
White Blood Cell Count	WBC	Leukocytes	BLOOD	HEMATOLOGY II		10 ⁹ /L	1	0.01
White Blood Cell Count	WBC	Leukocytes	BLOOD	HEMATOLOGY II		10 ⁹ /L	1	0.01
White Blood Cell Count	WBC	Leukocytes	BLOOD	HEMATOLOGY II	10*3	10 ⁹ /L	1	0.01
White Blood Cell Count	WBC	Leukocytes	BLOOD	HEMATOLOGY II	10*3/mm3	10 ⁹ /L	1	0.01
White Blood Cell Count	WBC	Leukocytes	BLOOD	HEMATOLOGY II	10*3/uL	10 ⁹ /L	1	0.01
White Blood Cell Count	WBC	Leukocytes	BLOOD	HEMATOLOGY II	10*9/L	10 ⁹ /L	1	0.01
White Blood Cell Count	WBC	Leukocytes	BLOOD	HEMATOLOGY II	U/L	10 ⁹ /L	1	0.01
White Blood Cell Count	WBC	Leukocytes	BLOOD	HEMATOLOGY II	cell/uL	10 ⁹ /L	0.001	0.01
white_blood_cell_count_wbc	WBC	Leukocytes	BLOOD	HEMATOLOGY II	10_3_mcl	10 ⁹ /L	1	0.01
white_blood_cell_count_wbc	WBC	Leukocytes	BLOOD	HEMATOLOGY II	k_ul	10 ⁹ /L	1	0.01
white_blood_cell_count_wbc	WBC	Leukocytes	BLOOD	HEMATOLOGY II	unknown_unit	10 ⁹ /L	1	0.01
white_blood_cell_count_wbc	WBC	Leukocytes	BLOOD	HEMATOLOGY II	x10_3_uL	10 ⁹ /L	1	0.01

Integration of Baseline Characteristics

- Develop a consolidated baseline domain to capture the baseline characteristics, such as diagnosis, ECOG, disease stage, biomarker, etc.. These assessments are typically recorded only once and is represented in different values/formats in RWD. Distributing them across multiple domains adds avoidable complexity, especially at the stage of ADaM derivation.

ZCTESTCD	Preamble	Flatiron	Connect MM	COTA	Standard Value
DEL17P	N	ABSENT	N	NEGATIVE	ABSENT
DEL17P	Y	PRESENT	Y	POSITIVE	PRESENT
ISS	STAGE I	STAGE I		I	STAGE I
ISS	STAGE II	STAGE II		II	STAGE II
ISS	STAGE III	STAGE III		III	STAGE III
BNL	N	N		ABSENT	N
BNL	Y	Y		PRESENT	Y

Standardization of Treatment Regimens

- Treatment regimen names may include generic names, brand names, or free text. Develop an Excel file listing unique regimens and review it with clinicians and statisticians to maintain standardized naming conventions throughout the programming process.

Original Treatment Name	Standard Treatment Name after Adjudication
Revlimid(lenalidomide)	LENALIDOMIDE
Autologous SCT2	AUTOLOGOUS STEM CELL TRANSPLANT
CLINICAL STUDY DRUG	INVESTIGATIONAL DRUG (UNKNOWN)
Transplant	AUTOLOGOUS STEM CELL TRANSPLANT
	MISSING

Comorbidity Adjudication via Patient-Level File Review

- Applying comorbidity criteria using programming logic is not feasible due to the complexity of the data. Therefore, a patient-level file containing ICD-9 and ICD-10 codes, corresponding verbatim terms, and start and end dates will be extracted into an Excel file. Clinicians and statisticians will review this file to determine subject inclusion or exclusion based on each patient's comorbidity profile.

Original Comorbidity Term	Exclusion Criteria
Congestive heart failure	Myocardial infarction or other cardiac function-related disease (eg, unstable angina, congestive heart failure New York Heart Association Class III-IV), or pericardial disease.
SOLID TUMOR MALIGNANCY: MENINGIOMA	History of malignancies other than MM

Manual Adjudication of Line of Therapy

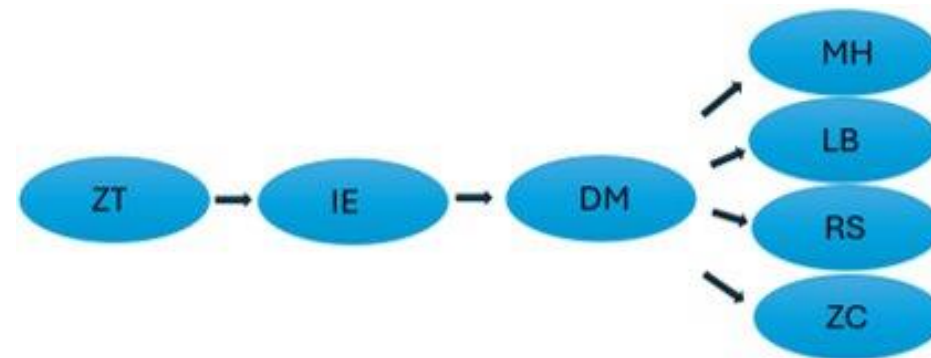
- An Excel file containing treatment start and end dates, corresponding outcomes, and relevant transplant information will be created by the programming group and provided to clinicians for manual adjudication. Following clinical adjudication, the following steps will be performed:
- Post-process line-of-therapy starting and ending dates
 - Deduplicate regimens and alphabetize component drugs within the same line of therapy

12Jun2008	Y	MAY20	1	[dexamethasone, lenalidomide]	.	30Jul2008	11Nov2008	PD PD PR PD PR PD	28SEP2017 15FEB2014 to 15NOV2017 23FEB2018 15DEC2018 to 15JUN2019 05NOV2019 06NOV2020	1	[dexamethasone, lenalidomide], [melphalan], [autologous SCT], [dexamethasone, lenalidomide]	30Jul2008	15Nov2017	
			2	[melphalan], [autologous SCT], [dexamethasone, lenalidomide]	progression	.	15Dec2008			15Nov2017	1	[dexamethasone, lenalidomide], [melphalan], [autologous SCT], [dexamethasone, lenalidomide]	30Jul2008	15Nov2017
			3	[bortezomib, cyclophosphamide, dexamethasone], [melphalan], [autologous SCT], [lenalidomide]	progression	.	20Dec2017			15Jun2019	2	[bortezomib, cyclophosphamide, dexamethasone], [autologous SCT], [lenalidomide]	20Dec2017	15Jun2019
			4	[daratumumab, dexamethasone, pomalidomide]	toxicity	.	25Jul2019			15Feb2020	3	[daratumumab, dexamethasone, pomalidomide]	25Jul2019	15Feb2020
			5	[carfilzomib, dexamethasone]	unknown_due_to_missing_documentation	.	8Dec2020			15Feb2021	4	[carfilzomib, dexamethasone]	8Dec2020	15Feb2021

SDTM Domain Creation Order for RWD

— Unlike clinical trial workflows, RWE studies in Oncology/Hematology should begin with line-of-therapy derivation where the qualified line or index date will be determined. Once line of therapy is established, other domains will be derived in following order:

- ZT (Line of Therapy)
- IE (Inclusion/Exclusion Criteria Not Met)
- DM (Demographics)
- LB (Laboratory Test Results)
- ZC (Baseline Characteristics)
- MH (Medical History)
- RS (Disease Response)



— The line of therapy-derived index date drives the determination of qualified subjects versus non-qualified subjects.

Strong Documentation Practices

— Maintain clear documentation of:

- Lab test selection and mapping
- Lab unit conversion table
- Regimen mapping harmonization

LAB_UNIT_CONV LAB_MAP TRT_CONV PT_TERM_CONV TRT_CLASS_CONV

- Line-of-therapy adjudication file

 connectmm-adjudicated-lot-final

 cota-adjudicated-lot-final

 flatiron-adjudicated-lot-final

 Preamble-adjudicated-lot-final

- Subject-level comorbidity file

 Comorbidity_exclude_list_100125

Explicit Missing Data Imputation Rules

- Include detailed date imputation rules in the SAP for:
 - Partial dates
 - Death date imputation
 - Regimen start and end date logic
- Document any rules not covered by the SAP within the SDTM specifications.

SDTM IG Compliance with Necessary Flexibility

- Follow the SDTM IG as closely as possible, while introducing pragmatic flexibilities when required. Examples include:
 - Mapping disease diagnosis date to SUPPDM
 - Mapping ICD-9/10 codes to SUPPMH when no standard MH variable exists
 - Storing MHTERM text exceeding 200 characters in SUPPMH
 - Consolidating baseline characteristics into a single customized domain (ZC) rather than creating multiple small domains to improve programming efficiency

Consistent SEQ Logic Across Sources

- For BDS domains, apply identical sorting rules across all data sources to ensure consistent -SEQ variable derivation.

All the BDS domains need -seq:

RS:

- COTA: STUDYID, USUBJID, RSSPID, RSDTC, RSREFID, RSTESTCD
- Flatiron: STUDYID, USUBJID, RSSPID, RSDTC, RSREFID, RSTESTCD, RSORRES
- Preamble: STUDYID, USUBJID, RSSPID, RSDTC, RSREFID, RSTESTCD, RSORRES
- Connect-MM: STUDYID USUBJID, RSSPID, RSDTC, RSREFID, RSTESTCD, RSORRES

Suggested sorting order: STUDYID, USUBJID, RSTESTCD, RSDTC, RSREFID, RSORRES, RSSTRESC

ZC:

- COTA: STUDYID, USUBJID, ZCSPID, ZCREFID, ZCTESTCD, ZCDTC, ZCORRES, ZCORRESU
- Flatiron: STUDYID, USUBJID, ZCSPID, ZCREFID, ZCTESTCD, ZCDTC, ZCMETHOD, ZCORRES, ZCORRESU
- Preamble: STUDYID, USUBJID, ZCSPID, ZCREFID, ZCTESTCD, ZCDTC, ZCMETHOD, ZCORRES, ZCORRESU
- Connect-MM: STUDYID USUBJID SUBJID ZCTESTCD ZCSPID ZCDTC ZCORRES ZCSTRESC;

Suggested sorting order: STUDYID, USUBJID, ZCTESTCD, ZCDTC, ZCREFID, ZCMETHOD, ZCORRES, ZCORRESU, ZCSTRESC

ZT:

- COTA: STUDYID, USUBJID, ZTSPID, ZTGRPID
- Flatiron: STUDYID, USUBJID, ZTSPID, ZTGRPID
- Preamble: STUDYID, USUBJID, ZTSPID, ZTGRPID, ZTSTDTC, ZTENDTC, ZTTRT
- Connect-MM: STUDYID USUBJID ZTSPID ZTGRPID

Suggested sorting order: STUDYID, USUBJID, ZTGRPID

MH:

- COTA: STUDYID, USUBJID, MHSPID, MHTERM, MHCAT, MHSTDTC
- Flatiron: STUDYID, USUBJID, MHSPID, MHTERM, MHCAT, MHSTDTC
- Preamble: STUDYID, USUBJID, MHSPID, MHTERM, MHCAT, MHSTDTC, MHENDTC, MHREFID
- Connect-MM: STUDYID, USUBJID, MHSPID, MHTERM, MHCAT, MHSTDTC, MHENDTC, MHREFID

Suggested sorting order: STUDYID, USUBJID, MHREFID, MHTERM, MHCAT, MHSTDTC, MHENDTC

LB:

- COTA: STUDYID, USUBJID, LBSPID, LBTESTCD, LBSPEC, LBDTC, LBORRES, LBSTRESN, LBSTRESU, LBSTNRLO
- Flatiron: STUDYID, USUBJID, LBSPID, LBTESTCD, LBLOINC, LBSPEC, LBDTC, LBORRES, LBORNRL0, LBORNRI, LBSTRESN, LBSTRESU
- Preamble: STUDYID, USUBJID, LBSPID, LBTESTCD, LBSPEC, LBDTC, LBSTRESN, LBSTRESU
- Connect-MM: STUDYID USUBJID LBTESTCD LBSPID LBDTC LBSTRESN LBSTRESU LBORRES LBORRESU LBSPEC;

Suggested sorting order: STUDYID, USUBJID, LBTESTCD, LBDTC, LBREFID, LBORRES, LBSPEC, LBORNRL0, LBORNRI, LBSTRESN, LBSTRESU, LBLOINC

Early and Iterative P21 Validation

- Run Pinnacle 21 validation early to identify structural issues and critical warnings before ADaM development begins. Collaborate with statisticians and clinicians to resolve findings efficiently.

GLOBAL	TS0022	Missing VS dataset	1	0	High	Error
GLOBAL	SD1112	Missing TA dataset	1	0	Medium	Warning
GLOBAL	SD1113	Missing TE dataset	1	0	Medium	Warning
GLOBAL	SD1485	Missing LC domain when LB domain is present	1	0	Medium	Error
GLOBAL	TS0021	Missing CM dataset	1	0	Medium	Error
GLOBAL	TS0050	Missing PC dataset	1	0	Medium	Warning
GLOBAL	TS0051	Missing PP dataset	1	0	Medium	Warning

LB	SD1353	LBSTRESC does not equal LBORRES when LBSTRESU is equal to LBORRESU	725	0	Medium	Warning
LB	SD1470	LBORRES is not numeric value, when LBORNRLO is populated	1,231	0	Medium	Warning
LB	SD1471	LBORRES is not numeric value, when LBORNRHI is populated	1,200	0	Medium	Warning
LB	SD1472	LBORRES is not numeric value, when LBSTNRLO is populated	1,231	0	Medium	Warning

KEY LESSONS AND TAKEAWAYS FOR RWD-TO-SDTM WORKFLOWS

Clinical Collaboration

Engaging clinical experts early provides essential context for interpreting complex real-world data variables and prevents errors.

Process Scalability

Focusing on reusable processes and harmonization files supports consistent and scalable workflows across multiple projects.

Clear Documentation

Maintaining clear documentation and ensuring traceability is very important in the regulatory submission process.

Flexibility and Quality Validation

Flexible SDTM design combined with early, repeated quality checks ensures datasets meet regulatory and scientific standards.

Contact Information: jennypeishan.han@bms.com

Thank You!

