



Grünenthal
Think innovation.
Feel life.®

Don't Let Your Data Handle You

A Novel Approach to Clinical Programming

Michael S. Rimler (GSK)

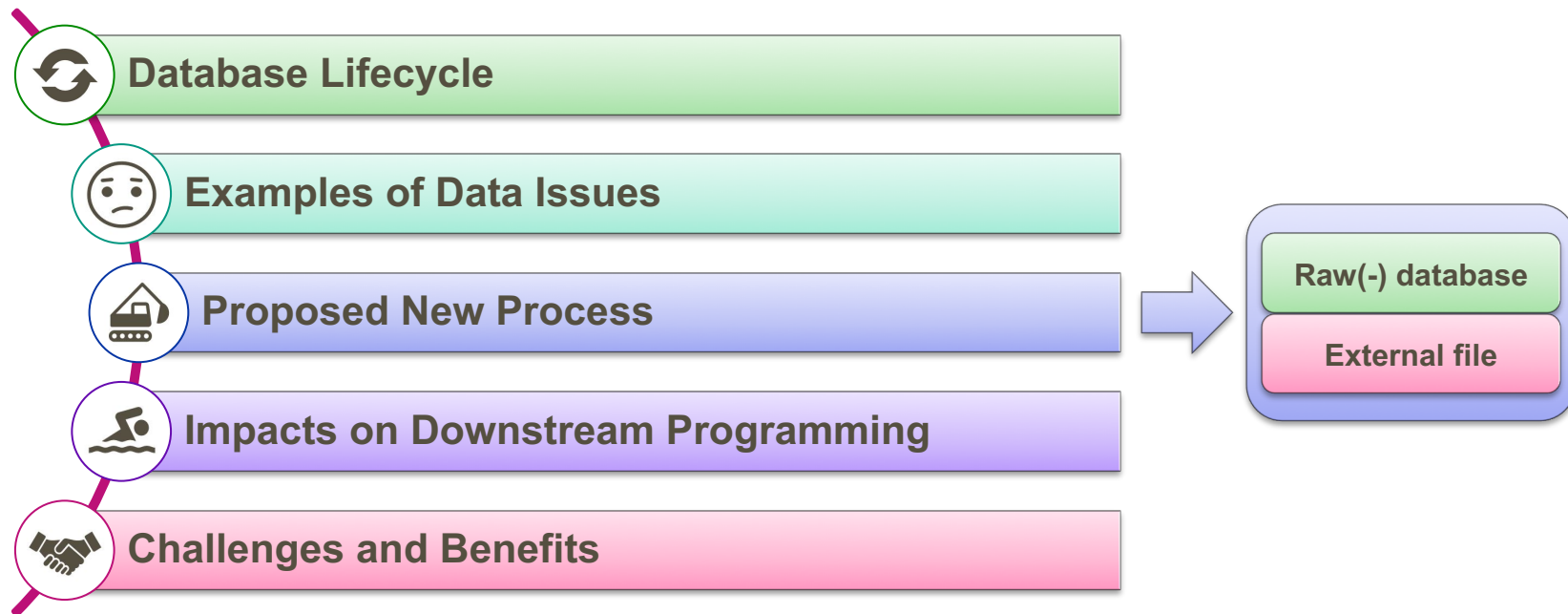
Jorine Putter (Grünenthal GmbH)

Issue at hand

- Arguably the biggest pain point for Statistics and Programming (S&P) groups working with dirty data prior to database lock (DBL)
 - Must ensure programs execute without error
 - Must illustrate that any anomalous results are due to dirty data and not issues with programming
- Data Management (DM) also spends effort managing data issues discovered by S&P

Proposed solution

- ‘Clean’ the raw data transferred from DM prior to mapping SDTM
 - Improves likelihood of error-free execution of downstream programming
 - Anomalous results are documented (semi-automatically)
 - Reduces effort required of DM with respect to S&P reported data issues



A large, stylized orange pill with a glossy, 3D effect, positioned on the left side of the slide. It has a gradient from a lighter orange at the top to a darker orange at the bottom, with a bright highlight on the upper left edge.

Clinical Database Lifecycle



Grünenthal
Think innovation.
Feel life.®

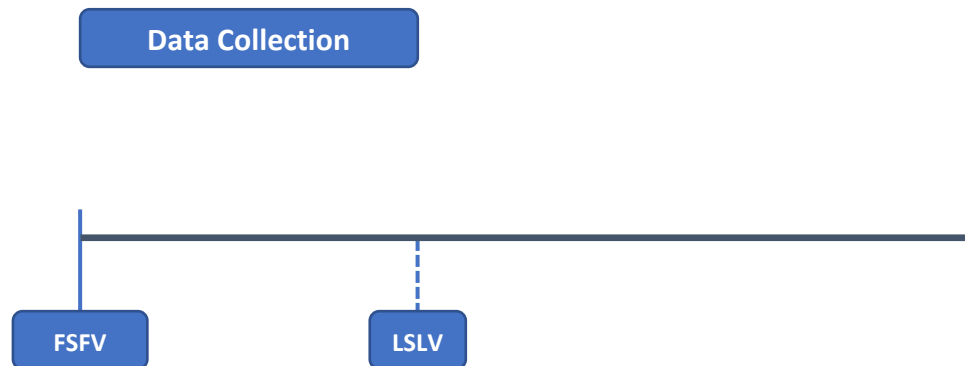
Clinical Database Lifecycle



Three Concurrent Phases

Data Collection

- Sites input patient data into the electronic case report form (eCRF)
- Newly collected data is at risk for entry error, even with automatic entry checks



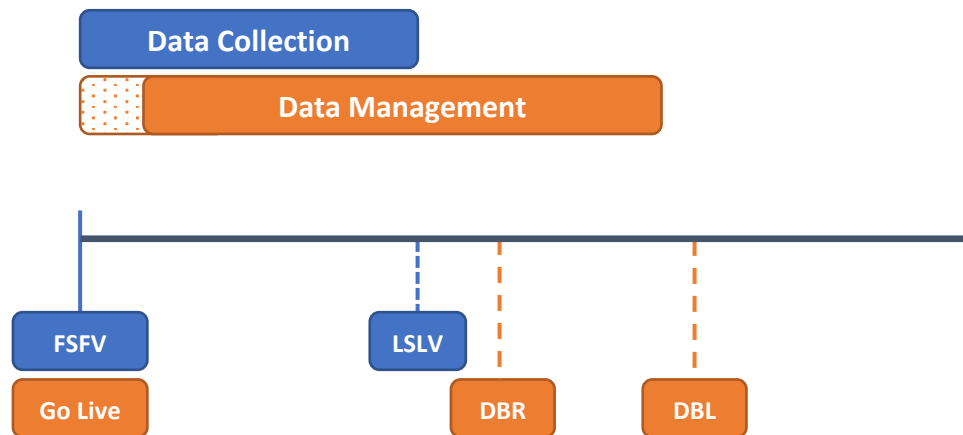
Clinical Database Lifecycle

Three Concurrent Phases

Data Collection

Data Management

- Review the collected data to ensure its completeness and quality
 - Automatic query generation embedded in the EDC system
 - Programmed edit checks
 - Reconciliation of externally provided results
 - Reviewing potential data issues flagged by Statistics and Programming (S&P) groups



Clinical Database Lifecycle

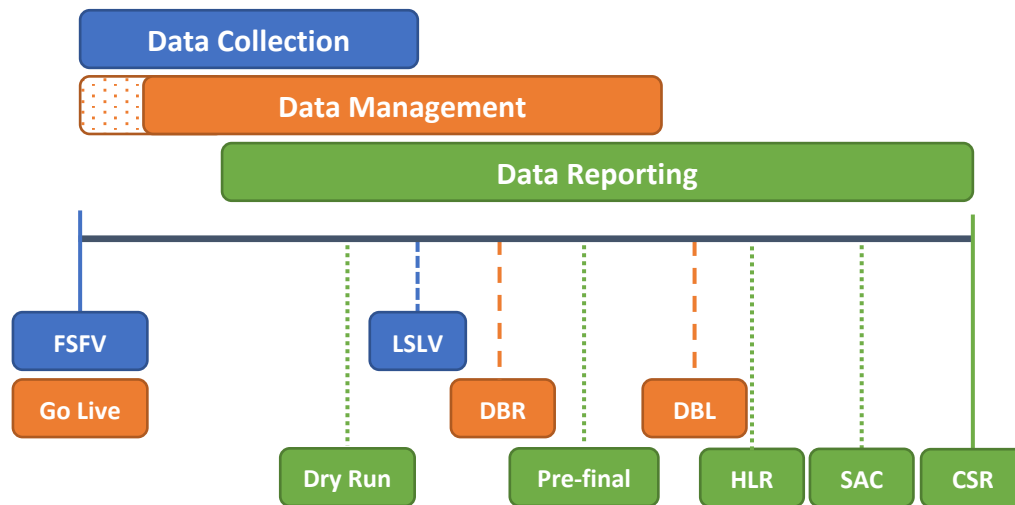
Three Concurrent Phases

Data Collection

Data Management

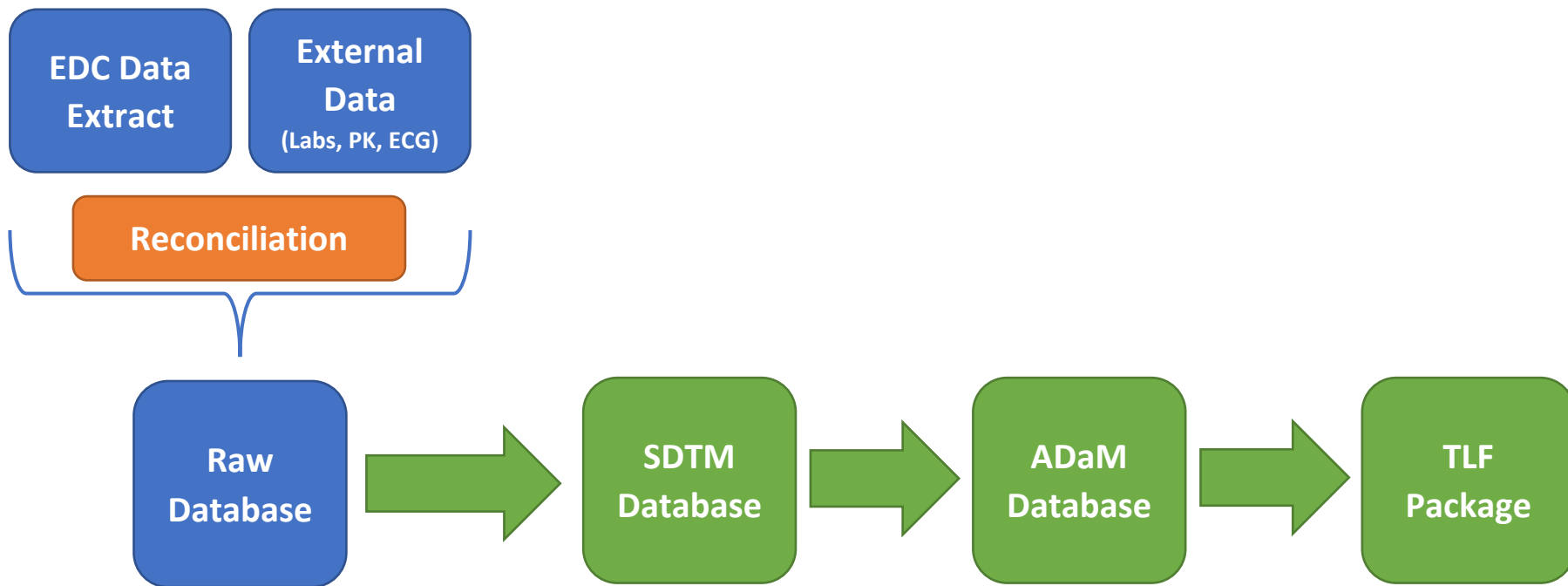
Data Reporting

- Ingest data as collected
- Process it, analyze it, and report the results in support of the analysis plan
- Creation of databases (SDTM and ADaM) and displays (TLFs)



Clinical Database Lifecycle

Standard Process





Examples of Data Issues



Grünenthal
Think innovation.
Feel life.®

Hypothetical Situation #1

Concomitant Medications Data

Patient Number	Medication	Medication Start Date	Medication Stop Date
1001	Medication 1	1-Jun-18	UNK
1001	Medication 2	14-Jun-18	13-Jun-18
1002	Medication 1	5-Jun-18	8-Jun-18
1002	Medication 2	7-Jun-18	xx-Jun-18
1003	Medication 2	18-Jun-18	21-Jun-18
1004	Medication 1	3-Jun-18	30-Jun-18
1004	Medication 1	3-Jun-18	30-Jun-18
1004	Medication 2	10-Jun-18	14-Jun-18
1005	Medication 2	8-Jun-18	Ongoing
1005	Medication 1	4-Jun-18	4-Jun-18

Should be flagged to DM for query

Issue #1

- Patient 1001, Medication 2: Medication Stop Date precedes Medication Start Date

Issue #2

- Patient 1004 has seemingly duplicate record (verbatim term happens to be misspelled)

Hypothetical Situation #2

Adverse Events and Disposition Data

Patient Number	AE Term	AE Start Date	AE Stop Date
2002	Fatigue	1-Jul-18	8-Jul-18
2003	Dizziness	10-Jul-18	20-Jul-18
2003	Stroke	20-Jul-18	20-Jul-18
2003	Death	18-Jul-18	20-Jul-18
2004	Nausea	1-Jul-18	5-Jul-18
2004	Fatigue	1-Jul-18	5-Jul-18

Patient Number	Discontinuation Date	Death Date
2001	-	-
2002	-	-
2003	19-Jul-18	18-Jul-18
2004	-	-
2005	-	-

Issue (Patient 2003)

- AE of Death has duration of 3 days (18-Jul to 20-Jul)
 - AE Start Date precedes AE Start Date for Stroke
- Multiple AE Stop Dates after Discontinuation Date (19-Jul)
- Death Date from Disposition data recorded as 18-Jul

**All dates should be flagged to DM
for query!!!**

Hypothetical Situation #3

Laboratory and Exposure Data

Exposure Module		
Patient Number	EX Start Date	Ex Start Time
3001	1-Jul-18	9:00
3002	10-Jul-18	9:15
3003	20-Jul-18	8:55
3004	18-Jul-18	9:02
3005	1-Jul-18	8:47
3006	1-Jul-18	9:01

Lab Module		
Patient Number	Lab Start Date	Lab Start Time
3001	1-Jul-18	8:58
3002	10-Jul-18	9:12
3003	20-Jul-18	8:50
3004	18-Jul-18	9:03
3005	1-Jul-18	8:30
3006	1-Jul-18	9:01

Issue #1

- Patient 3004 has Lab Draw after first dose date/time. Is this correct?

**Should be flagged
to DM for query**

Issue #2

- Patient 3006 has Lab Draw at exact same date/time as first dose. Was it administered pre-dose?

Hypothetical Situation #4

Vital Sign Data

Patient Number	Systolic Blood Pressure	Diastolic Blood Pressure	Height (cm)	Weight (kg)
4001	128	80	165	75
4002	125	78	140	85
4003	130	82	6	220
4004	80	120	185	92
4005	120	90	172	83

Issue #1

- Patient 4002 height and weight appear disproportionate for typical patient

Issue #2

- Patient 4003 height and weight appear measured in incorrect units

Issue #3

- Patient 4004 systolic and diastolic blood pressure measurement appear entered in reverse

**Should be flagged
to DM for query**

What are we illustrating?

- Prior to DBL, data transfers from DM to S&P will contain **dirty data**
 - Even with automatic and programmed edit checks by DM
 - Even though DM data cleaning is ongoing
 - Even if data cut off dates are applied
- S&P often must **revise code** (SDTM, ADaM, and/or TLF) to accommodate unexpected dirty data

Our proposal

- Aims to mitigate the impact of reporting with dirty data
- Extra benefits:
 - Improved communication and tracking of data issues flagged by S&P
 - Improved mechanism for handling data issues that persist after DBL

A large, stylized orange pill with a glossy, 3D effect, positioned on the left side of the slide. It has a gradient from a lighter orange at the top to a darker orange at the bottom, with a bright highlight on the upper right edge.

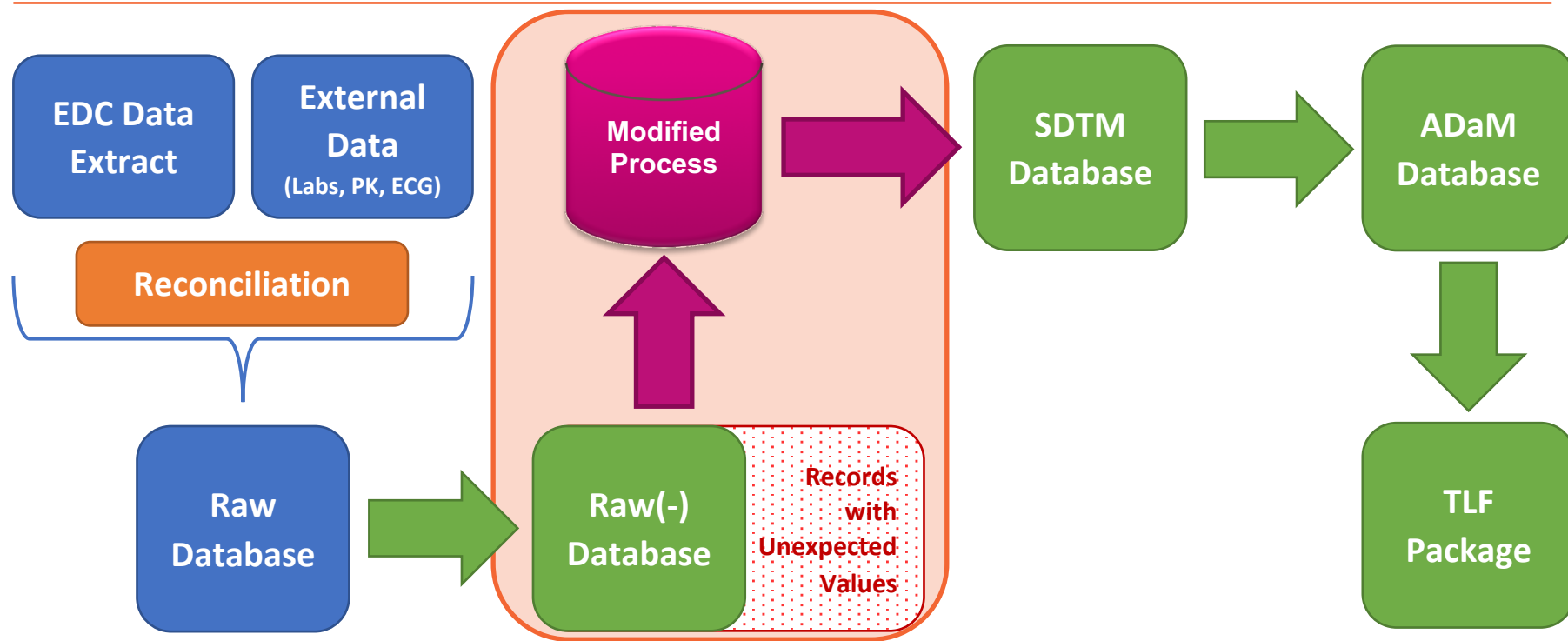
Proposed New Process



Grünenthal
Think innovation.
Feel life.®

Clinical Database Lifecycle

Proposed Process



Raw Data Processing



What is this Raw(-) Database?

Owned and managed by S&P (or appropriate group)

- Identifies (and separates) records with unexpected values
- Output is two mutually exclusive and exhaustive sets of data
 - Raw(-) database: Clean of any S&P defined data issues
 - External file: Dirty data ready for potential data query

Handling of Records with Unexpected Values

- Compiled into an external database file (e.g., Microsoft Excel ®)
- File serves three purposes
 - ✓ Running log of all S&P flagged raw data issues throughout the study
 - ✓ Track DM data cleaning process at a high level
 - ✓ Mechanism for handling data issues that persist after DBL

External File Containing Unexpected Values

Spreadsheet with a sheet for each raw domain

- AE, CONMED, DEMO, LAB, etc.

One record for each field with an unexpected value

Patient Number	Systolic Blood Pressure	Diastolic Blood Pressure	Height (cm)	Weight (kg)
4001	128	80	165	75
4002	125	78	140	85
4003	130	82	6	220
4004	80	120	185	92
4005	120	90	172	83

Additional variable identifying the problematic field

PATID	SYSBP	DIABP	HGT	WGT	RawVar
4002	125	78	140	85	HGT
4002	125	78	140	85	WGT
4003	130	82	6	220	HGT
4003	130	82	6	220	WGT
4004	80	120	185	92	SYSBP
4004	80	120	185	92	DIABP

External File Containing Unexpected Values

Two additional variables

- Updated value used for programmatic handling of problematic records
 - Agreed upon by study team
 - Documented if persists after DBL
- Comment field for documentation purposes (e.g., can append this file to cSDRG)

Patient Number	Systolic Blood Pressure	Diastolic Blood Pressure	Height (cm)	Weight (kg)
4001	128	80	165	75
4002	125	78	140	85
4003	130	82	6	220
4004	80	120	185	92
4005	120	90	172	83

PATID	SYSBP	DIABP	HGT	WGT	RawVar	UpdVal	Comment
4002	125	78	140	85	HGT	170	xxxx
4002	125	78	140	85	WGT		
4003	130	82	6	220	HGT	183	xxxx
4003	130	82	6	220	WGT	100	xxxx
4004	80	120	185	92	SYSBP	120	xxxx
4004	80	120	185	92	DIABP	80	xxxx

External File Containing Unexpected Values



Three variables for Data Query Tracking

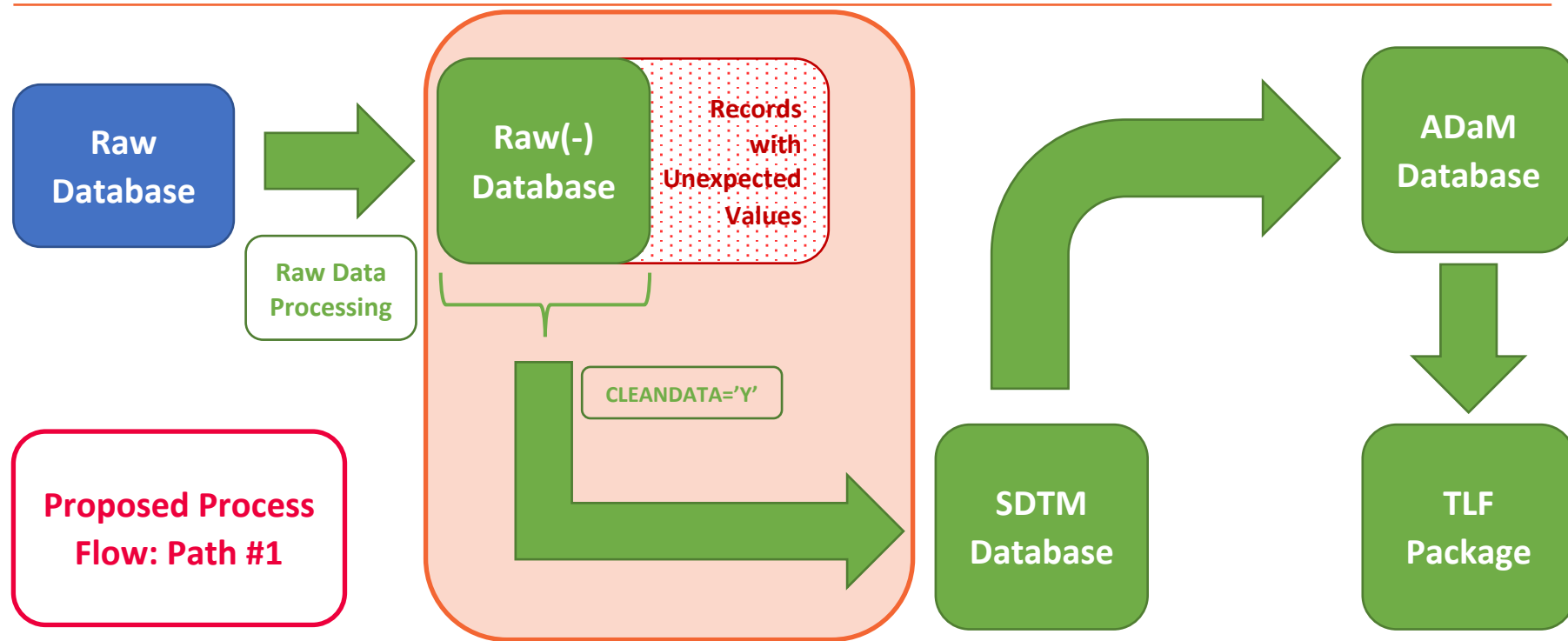
- **PRVFLAG**: Flag records if new or old
- **RSLVD**: For previously flagged issues, do they still exist (programmatic flag)?
- **QRY**: Did DM issue query?
 - Manually populated by DM

Patient Number	Systolic Blood Pressure	Diastolic Blood Pressure	Height (cm)	Weight (kg)
4001	128	80	165	75
4002	125	78	140	85
4003	130	82	6	220
4004	80	120	185	92
4005	120	90	172	83

PATID	SYSBP	DIABP	HGT	WGT	RawVar	UpdVal	Comment	PRVFLAG	RSLVD	QRY
4002	125	78	140	85	HGT	170	xxxx	Y	N	N
4002	125	78	140	85	WGT			Y	N	N
4003	130	82	6	220	HGT	183	xxxx	Y	N	Y
4003	130	82	6	220	WGT	100	xxxx	Y	N	Y
4004	80	120	185	92	SYSBP	120	xxxx	N		
4004	80	120	185	92	DIABP	80	xxxx	N		

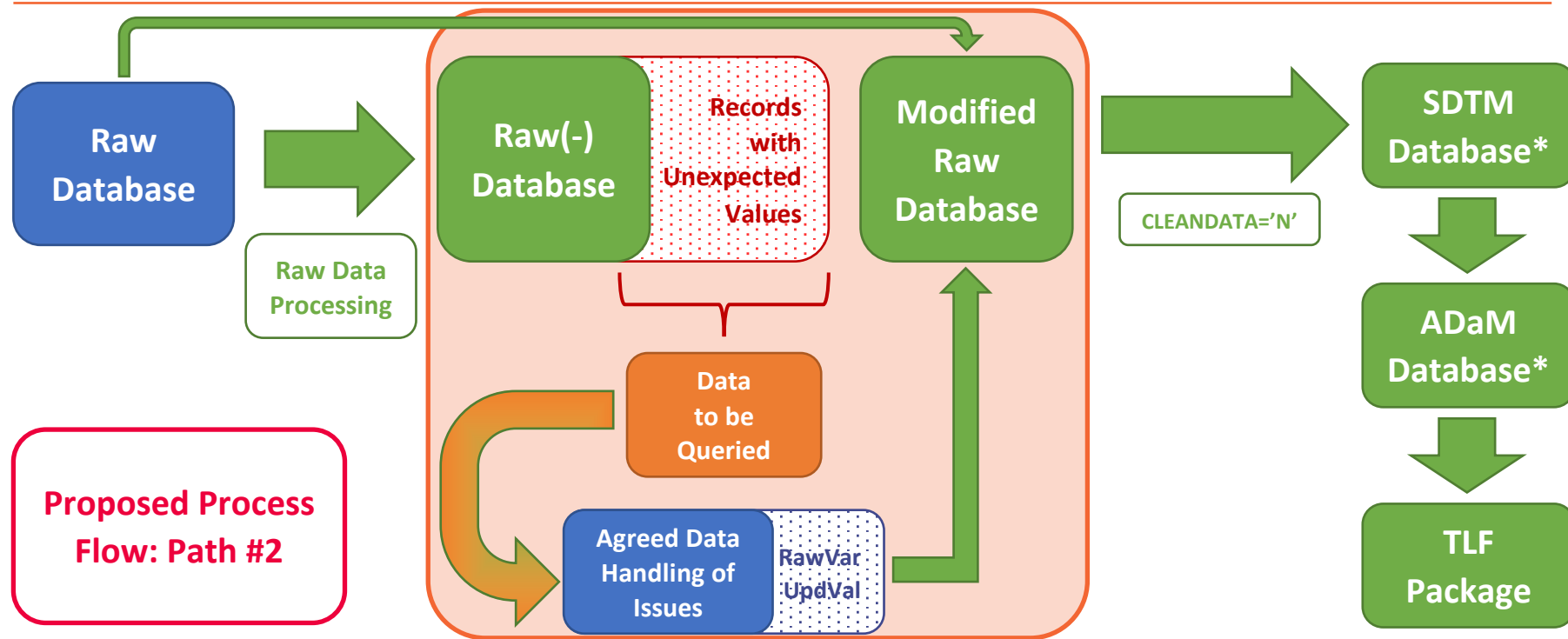
Proposed Process

Mapping SDTM using only clean data



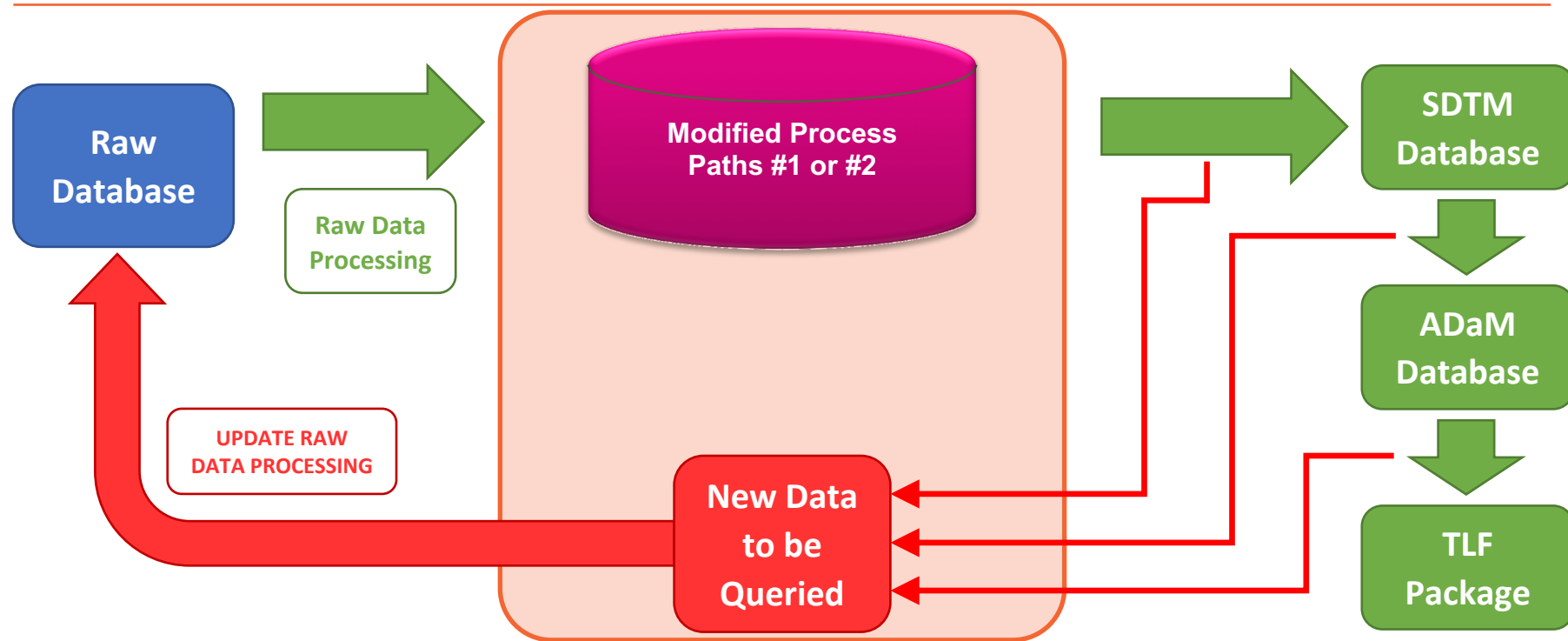
Proposed Process

Mapping SDTM using full database



Proposed Process

Feedback Mechanisms





Impacts on Downstream Programming



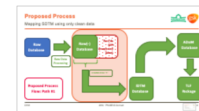
Grünenthal
Think innovation.
Feel life.®

Impact on SDTM Programming



CLEANDATA="Y" (Path #1)

- Clean data only, defined by S&P
- SDTM programming should execute without error (in theory)



CLEANDATA="N" (Path #2)

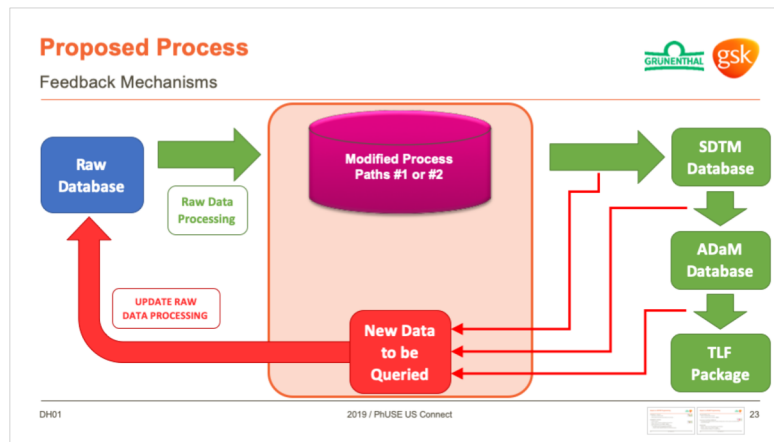
- Complete database
- Merges originally collected raw database with external file
- Modifies problematic records (**RawVar**, **UpdVal**)
- Utilize supplemental domains to track any modifications
 - Create SUPP record for any field appearing in the external file
 - If values are updated, create SUPP record to retain originally collected value



Impact on SDTM Programming

CLEANDATA in (“Y” “N”)

- If new issues found, update upstream raw data processing code to filter the record to the external file and rerun

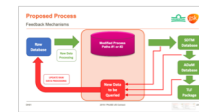


Input data (SDTM) is clean

- Path #1: Unexpected data is removed
- Path #2: Unexpected data is handled via **UpdVal**

If new issues are identified in ADaM code

- If new issues found, update upstream raw data processing code to filter the record to the external file and rerun SDTM and ADaM



Analysis Flags

- ANL98FL: Flag any record which appears in the external file
- ANL99FL: Flag any record modified by **UpdVal**
- Provides downstream flexibility in TLF programming

Challenges and Benefits



Grünenthal
Think innovation.
Feel life.®

The Challenges



Functional Area Pain Points

Statistics and Programming

- Data issues are arguably the biggest pain point
- Accountable to find data issues? Usually not explicitly
- However, to achieve a delivery milestone, S&P teams must have
 - Error free execution
 - 100% clean validation of all datasets and displays
 - Transparent lineage of each raw collected value to an analysis result

**S&P must be able to illustrate
that an anomalous analysis result
is due to dirty data and not a
programming issue**

Data Management

- Spending double effort to manage the data issues flagged by S&P
- Reconciling these issues with standard query tracking processes

The Challenges



Cultural Shift in Programming Approach

Additional Programming Step

- S&P teams invest time in upstream programming (raw data processing)
 - Facilitates smoother downstream programming
 - Improved data issue management mechanism

Task Ownership

- Allocating responsibility between S&P and DM groups
- Allocating responsibility across multiple organizations (e.g., sponsor and multiple CROs)

Communication Challenges

- Requires streamlined communication channels
- Example: New data extract
 - Requires an agreed remedy for unexpected data issue resolution

External File of Data Issues from S&P

- Unexpected data issues conveniently packaged for DM
 - Can be annotated to ensure queries are not duplicated for the remainder of the trial
 - As a running log, burden of DM review from S&P identified issues will ease during the study lifecycle
 - Allows for diagnostics such as number of open queries and number of queries resolved
 - Provides a vision of what the database will look like at DBL with respect to flagged issues (planning)

Improved downstream programming efficiency

- Developed on 'clean' raw data → Programmed to expected data
- Feedback mechanisms and emphasis placed on updating upstream S&P raw data processing
 - Reduces the need for downstream code maintenance

Alternative to Hardcoding

- External file allows for handling of persistent data issues post-DBL without hardcoding
 - For example, if a site is unresponsive or confirms a problematic data point is 'correct'
- Modifications of raw data is tracked and documented
 - Traceable in SDTM and ADaM
 - Transparent to regulatory
 - Include in cSDRG to facilitate review (Clinical Study Data Reviewer's Guide)

Michael S. Rimler
GlaxoSmithKline
michael.s.rimler@gsk.com

Jorine Putter
Grünenthal GmbH
Jorine.putter@grunenthal.com