GOETHE UNIVERSITÄT FRANKFURT AM MAIN
Institut für Digitale Medizin und Klinische Datenwissenschaften

# Chancen und Fallstricke der Real-World Datenanalyse

Janne Vehreschild



# Agenda

GOETHE UNIVERSITÄT FRANKFURT AM MAIN
Institut für Digitale Medizin und Klinische Datenwissenschaften

1. Why is everyone talking about real-world data these days?
2. What are structural differences between real-world data and data from prospective trials?
3. What are the chances in using real-world data for my research?
4. Which limitations and biases are connected to real-world data, and how should I encounter them in my analyses?

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

# 1. Why Real-World Data?

Patient

GEFÖRDERT VOM

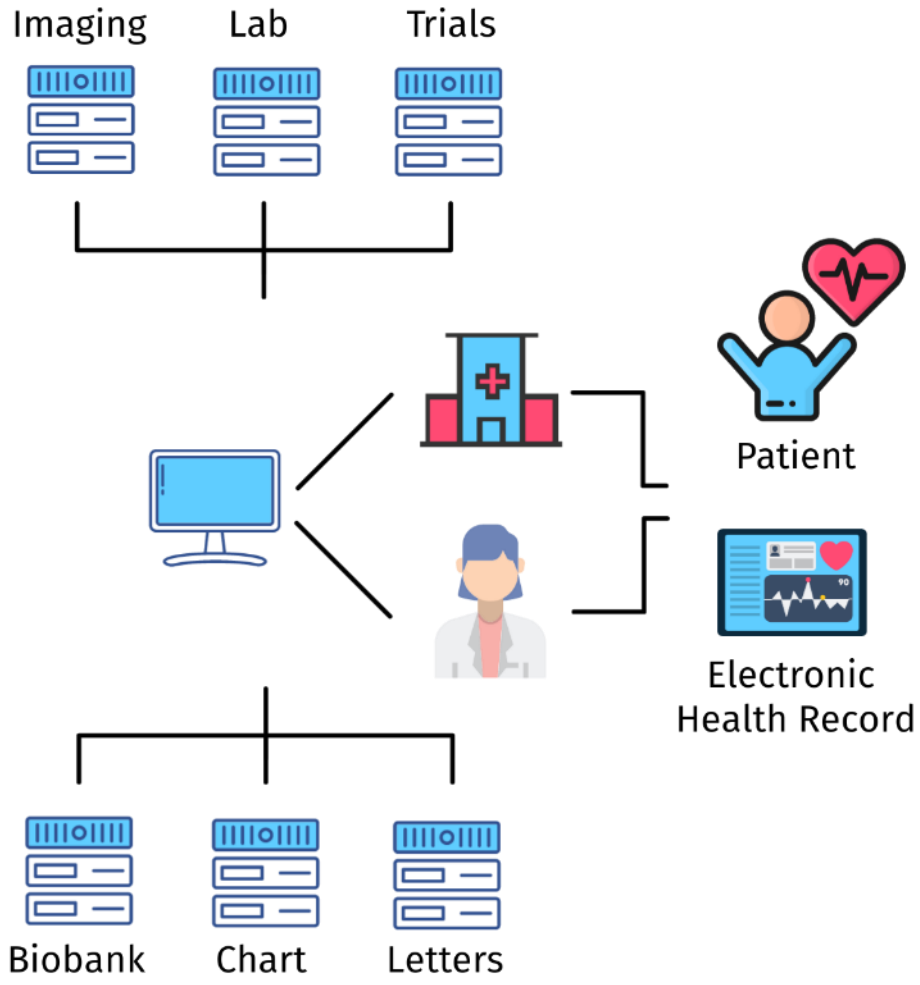
# 1. Why Real-World Data?

Patient

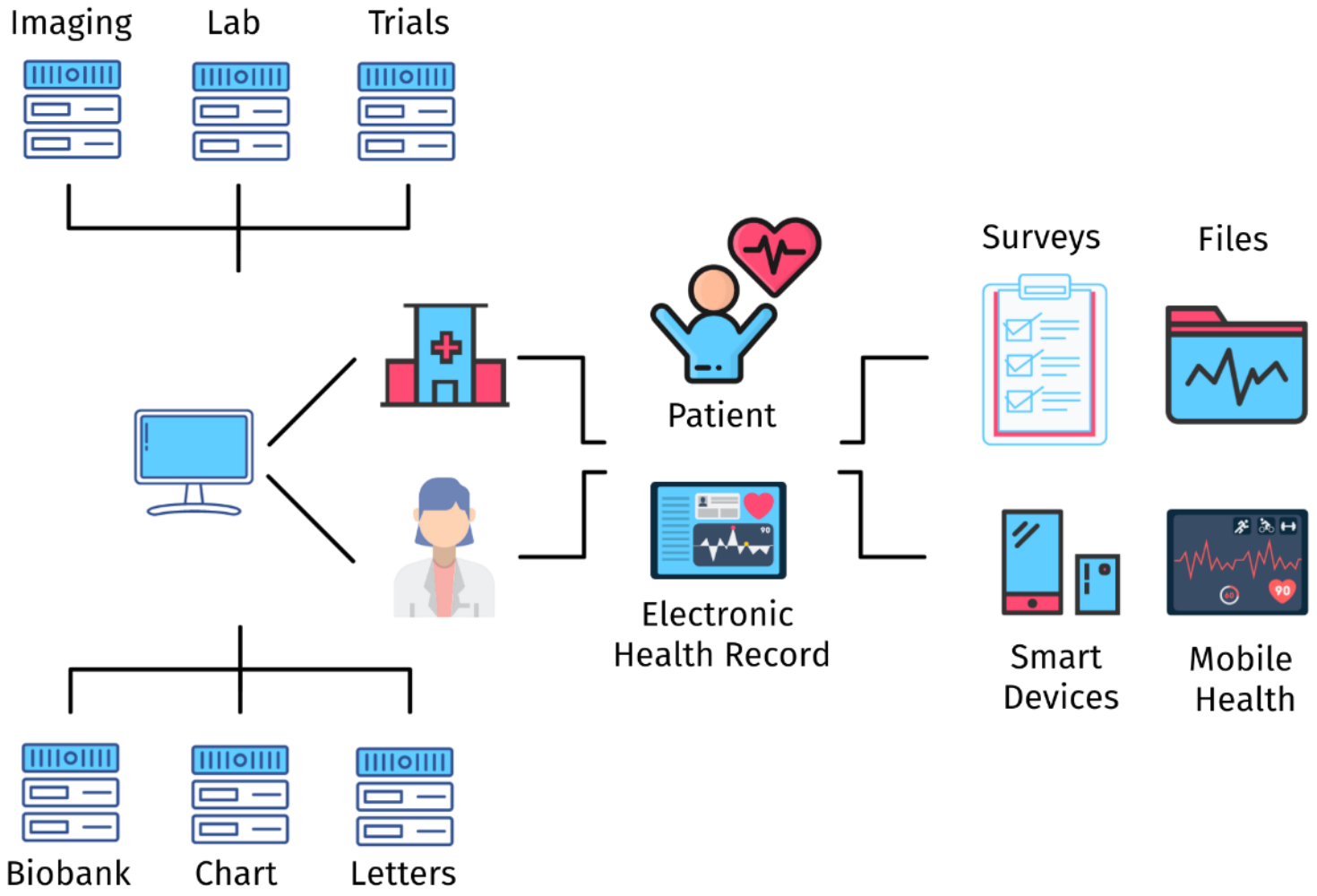
Electronic
Health Record

GEFÖRDERT VOM

# 1. Why Real-World Data?



# 1. Why Real-World Data?

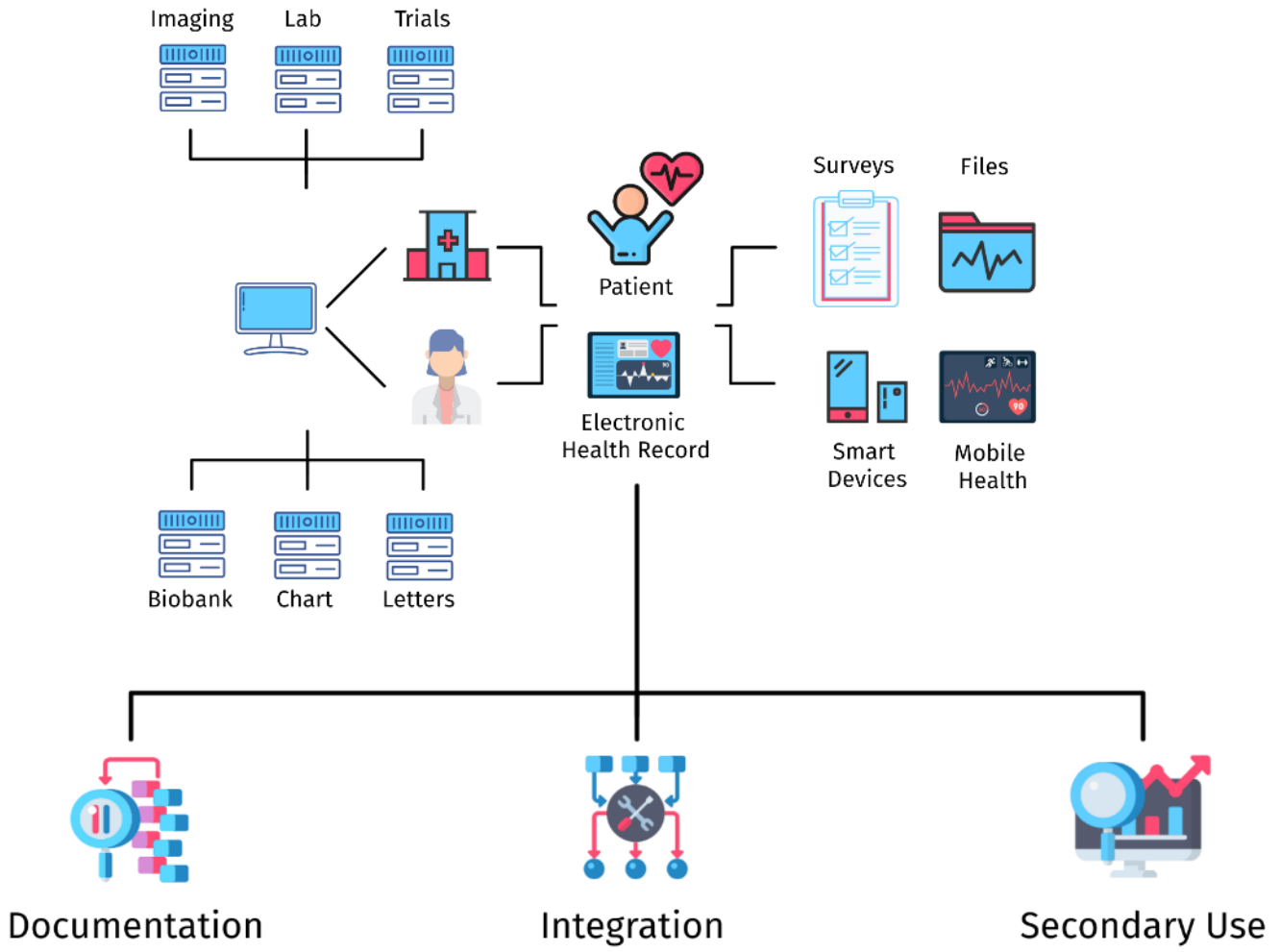



GOETHE UNIVERSITÄT FRANKFURT AM MAIN
Institut für Digitale Medizin und Klinische Datenwissenschaften

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

# 1. Why Real-World Data?

GOETHE UNIVERSITÄT FRANKFURT AM MAIN
Institut für Digitale Medizin und Klinische Datenwissenschaften




GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

# 2. Structural differences of data sources

Angelika

- 61 year old female
- No major comorbidities
- New diagnosis of right-sided colon cancer
- TNM: T3 N2 M0 (Stage III)
- Adjuvant therapy after successful resection
- **Study** treatment: FOLFOX followed by Pembrolizumab

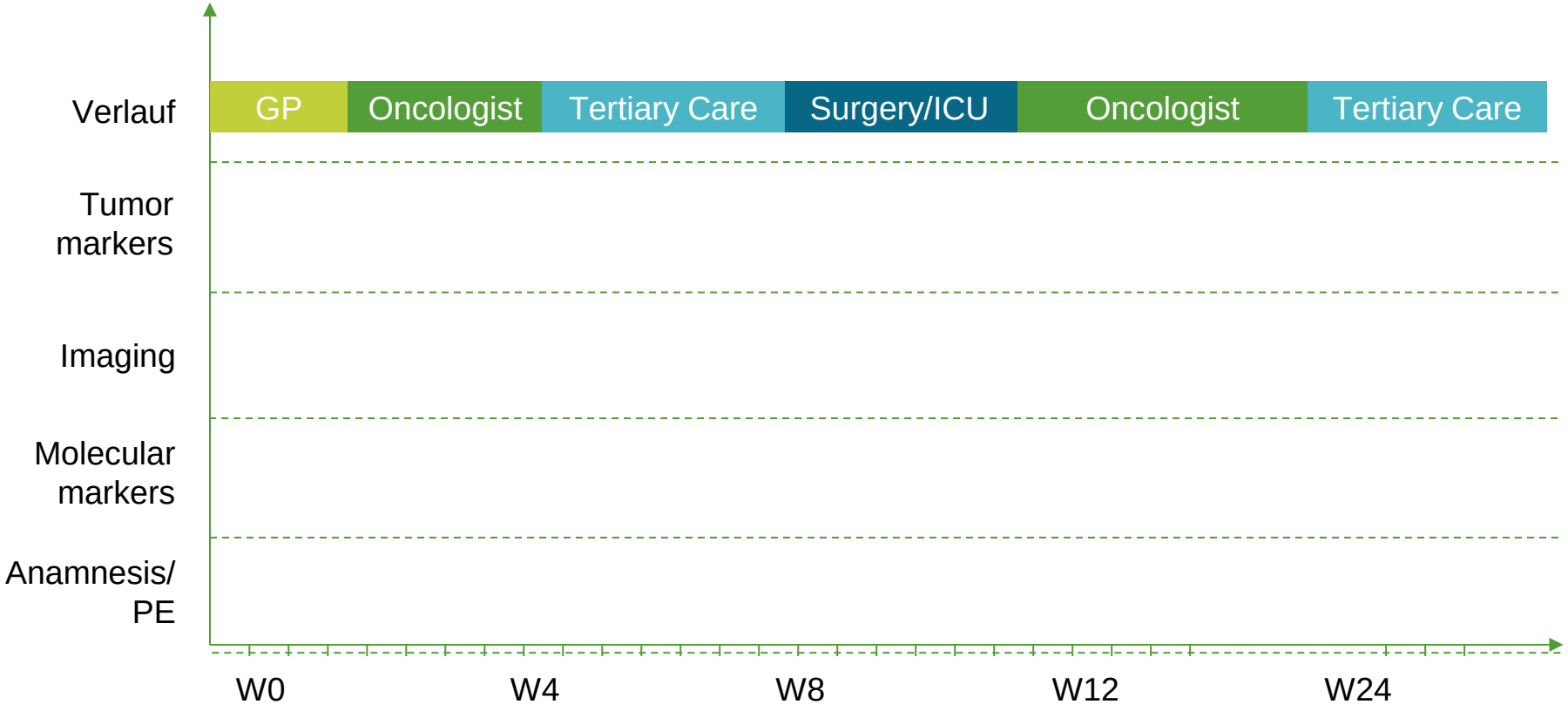
Horst

- 63 year old male
- No major comorbidities
- New diagnosis of right-sided colon cancer
- TNM: T3 N2 M0 (Stage III)
- Adjuvant therapy after successful resection
- **Standard** treatment: FOLFOX

GEFÖRDERT VOM

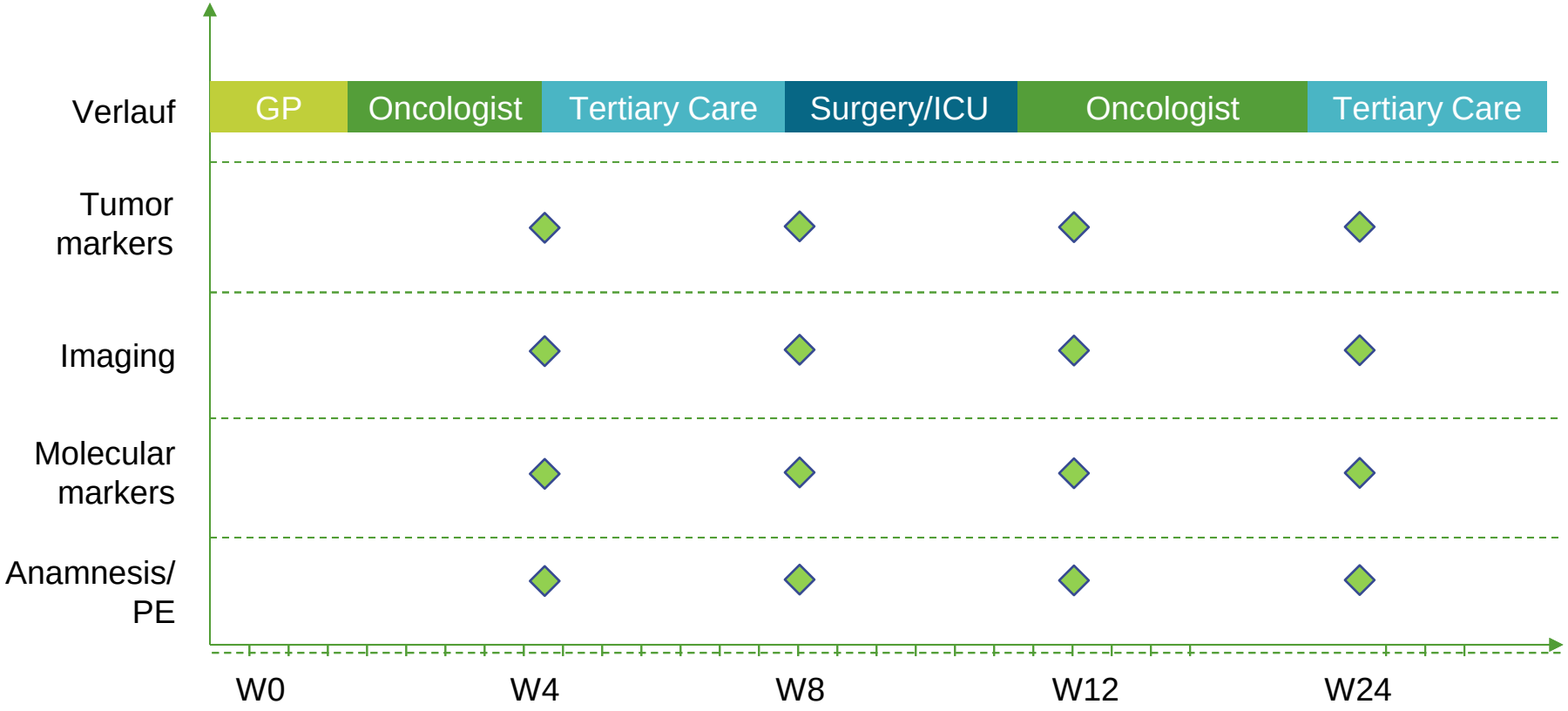
# 2. Structural differences of data sources

GOETHE UNIVERSITÄT FRANKFURT AM MAIN
Institut für Digitale Medizin und Klinische Datenwissenschaften




GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

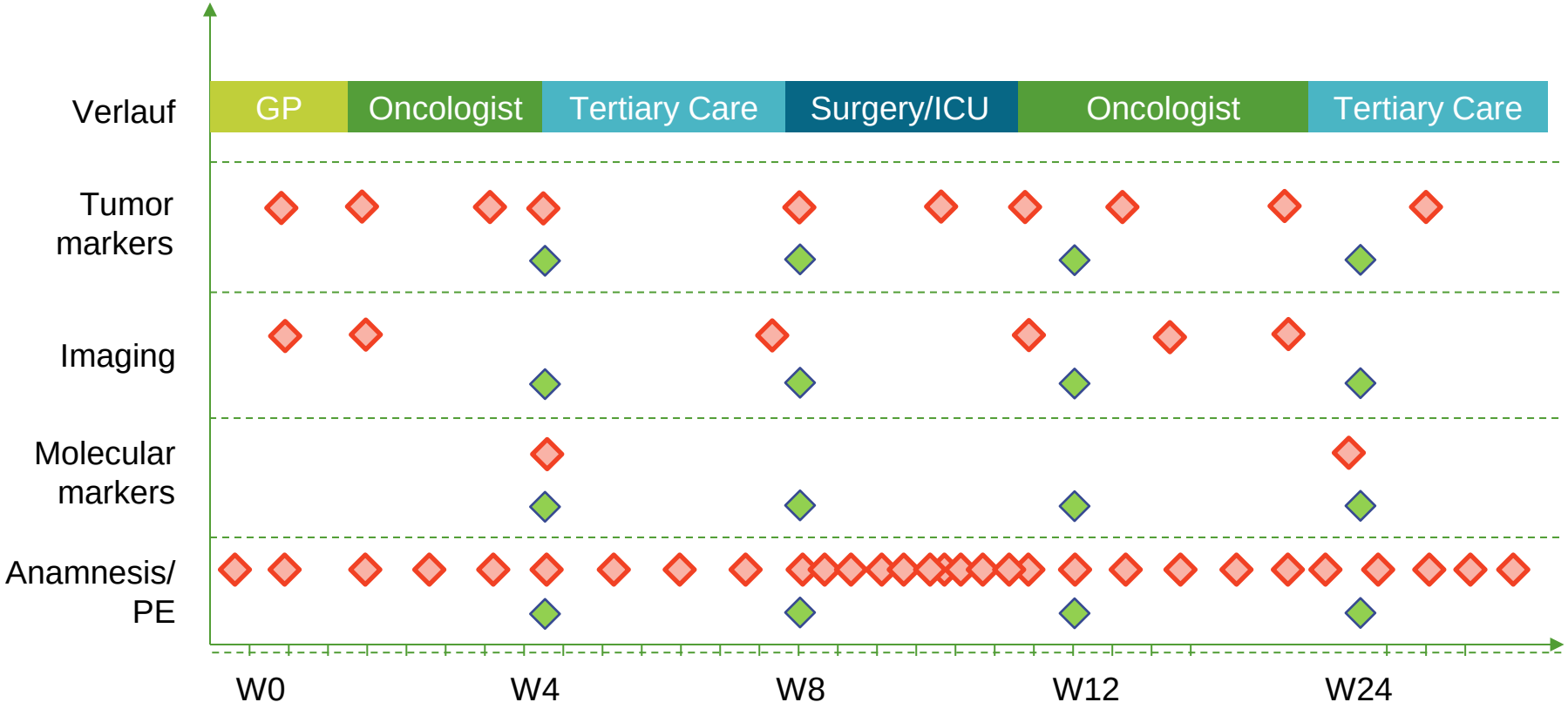
# 2. Structural differences of data sources




GEFÖRDERT VOM

# 2. Structural differences of data sources

GOETHE UNIVERSITÄT FRANKFURT AM MAIN
Institut für Digitale Medizin und Klinische Datenwissenschaften

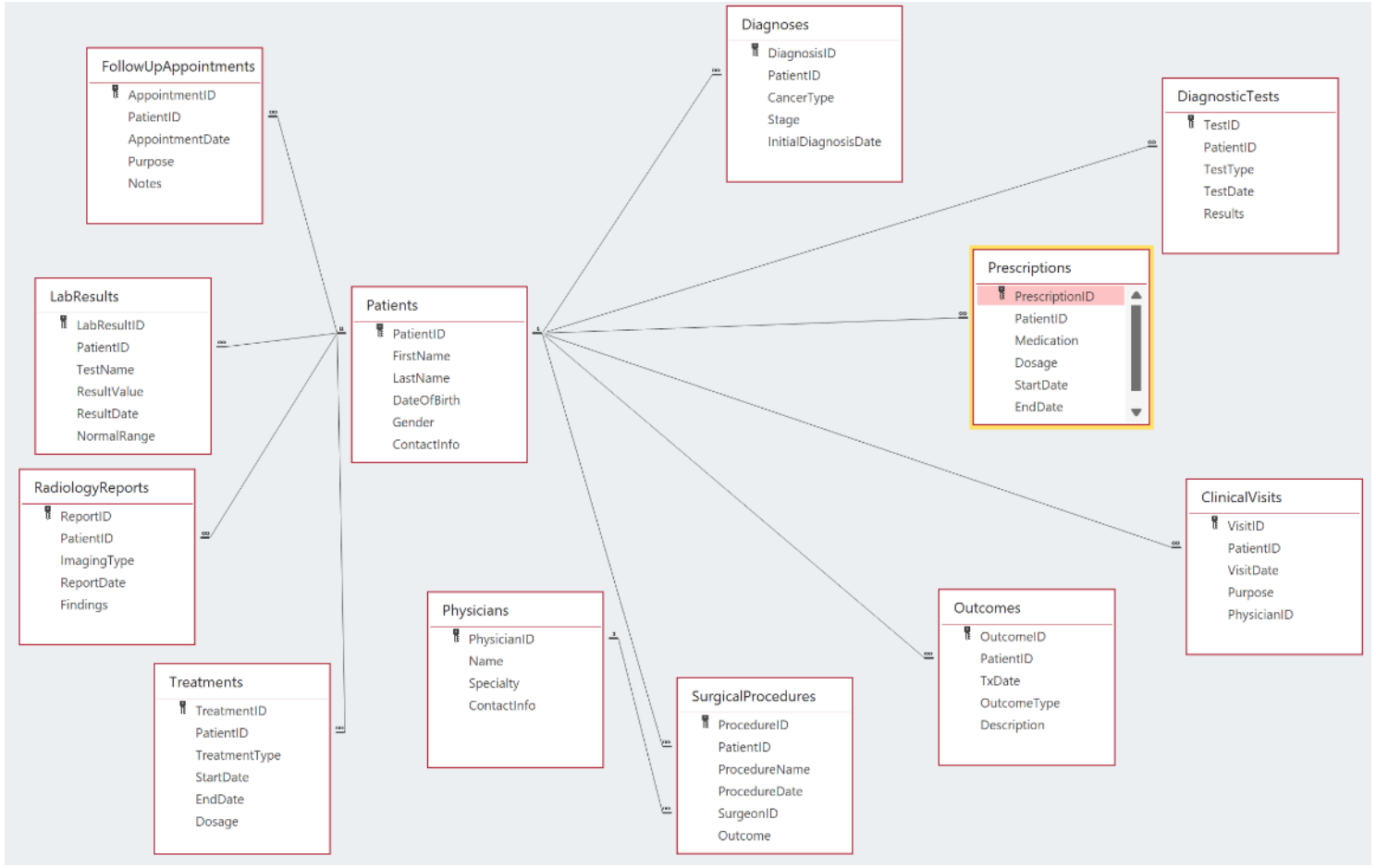



GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

# 2. Structural differences of data sources

| Patient ID | Treatment Group | Gender | Molecular Marker | Visit | Radiological Staging | Tumor Markers (CEA | LDH (U/L) | Physical Exam Results | Adverse Events |
|---|---|---|---|---|---|---|---|---|---|
| PAT-001 | Group A - Standard Therapy | Male | BRAF Mutant | Baseline | N0 | 40 | 147 | Abnormal | Mild |
| PAT-001 | Group A - Standard Therapy | Male | BRAF Mutant | Visit 1 | T3 | 40 | 217 | Normal | Mild |
| PAT-001 | Group A - Standard Therapy | Male | BRAF Mutant | Visit 2 | T3 | 10 | 167 | Abnormal | Moderate |
| PAT-001 | Group A - Standard Therapy | Male | BRAF Mutant | Visit 3 | N0 | 5 | 203 | Normal | Mild |
| PAT-001 | Group A - Standard Therapy | Male | BRAF Mutant | Visit 4 | M1 | 5 | 109 | Abnormal | Mild |
| PAT-002 | Group A - Standard Therapy | Female | BRAF Mutant | Baseline | M0 | 40 | 121 | Abnormal | Severe |
| PAT-002 | Group A - Standard Therapy | Female | BRAF Mutant | Visit 1 | T1 | 40 | 136 | Normal | Moderate |
| PAT-002 | Group A - Standard Therapy | Female | BRAF Mutant | Visit 2 | M1 | 40 | 187 | Normal | Moderate |
| PAT-002 | Group A - Standard Therapy | Female | BRAF Mutant | Visit 3 | N2 | 40 | 170 | Normal | Moderate |
| PAT-002 | Group A - Standard Therapy | Female | BRAF Mutant | Visit 4 | T1 | 40 | 188 | Normal | None |
| PAT-003 | Group A - Standard Therapy | Male | BRAF Mutant | Baseline | M1 | 30 | 240 | Normal | Mild |
| PAT-003 | Group A - Standard Therapy | Male | BRAF Mutant | Visit 1 | N2 | 40 | 158 | Abnormal | Mild |
| PAT-003 | Group A - Standard Therapy | Male | BRAF Mutant | Visit 2 | N0 | 10 | 139 | Abnormal | Mild |
| PAT-003 | Group A - Standard Therapy | Male | BRAF Mutant | Visit 3 | T4 | 30 | 187 | Normal | Moderate |
| PAT-003 | Group A - Standard Therapy | Male | BRAF Mutant | Visit 4 | N1 | 40 | 188 | Abnormal | None |
| PAT-004 | Group B - Experimental Therapy | Male | KRAS Mutant | Baseline | T3 | 20 | 181 | Abnormal | None |
| PAT-004 | Group B - Experimental Therapy | Male | KRAS Mutant | Visit 1 | T2 | 20 | 125 | Abnormal | Mild |
| PAT-004 | Group B - Experimental Therapy | Male | KRAS Mutant | Visit 2 | N0 | 10 | 177 | Normal | None |
| PAT-004 | Group B - Experimental Therapy | Male | KRAS Mutant | Visit 3 | N1 | 5 | 172 | Abnormal | None |
| PAT-004 | Group B - Experimental Therapy | Male | KRAS Mutant | Visit 4 | M1 | 30 | 109 | Abnormal | Mild |
| PAT-005 | Group A - Standard Therapy | Male | KRAS Wild-Type | Baseline | T1 | 20 | 248 | Normal | Severe |

# 2. Structural differences of data sources




GEFÖRDERT VOM

# 3. Chances of Real-World Data

GOETHE UNIVERSITÄT FRANKFURT AM MAIN
Institut für Digitale Medizin und Klinische Datenwissenschaften

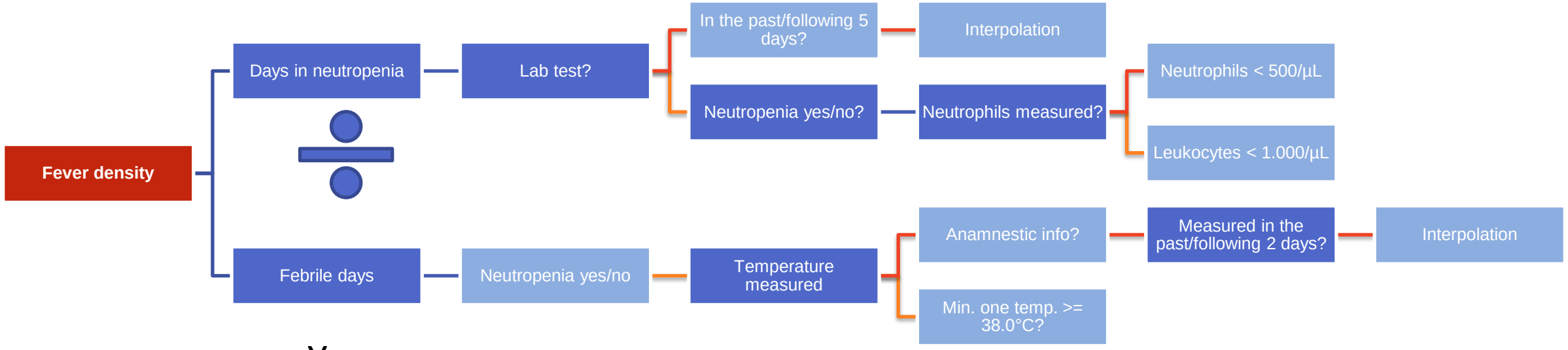
- Access all existing clinical knowledge and experience (hypothetically)
- Do so at minimum expense of time and resources (hypothetically)

- Use statistical power to:
  - Reveal hard to detect associations between clinical courses / decisions and outcome
  - Define more accurate disease phenotypes to instruct Omics-based research
  - Predict outcome and individualize strategies

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

# Complex disease phenotypes

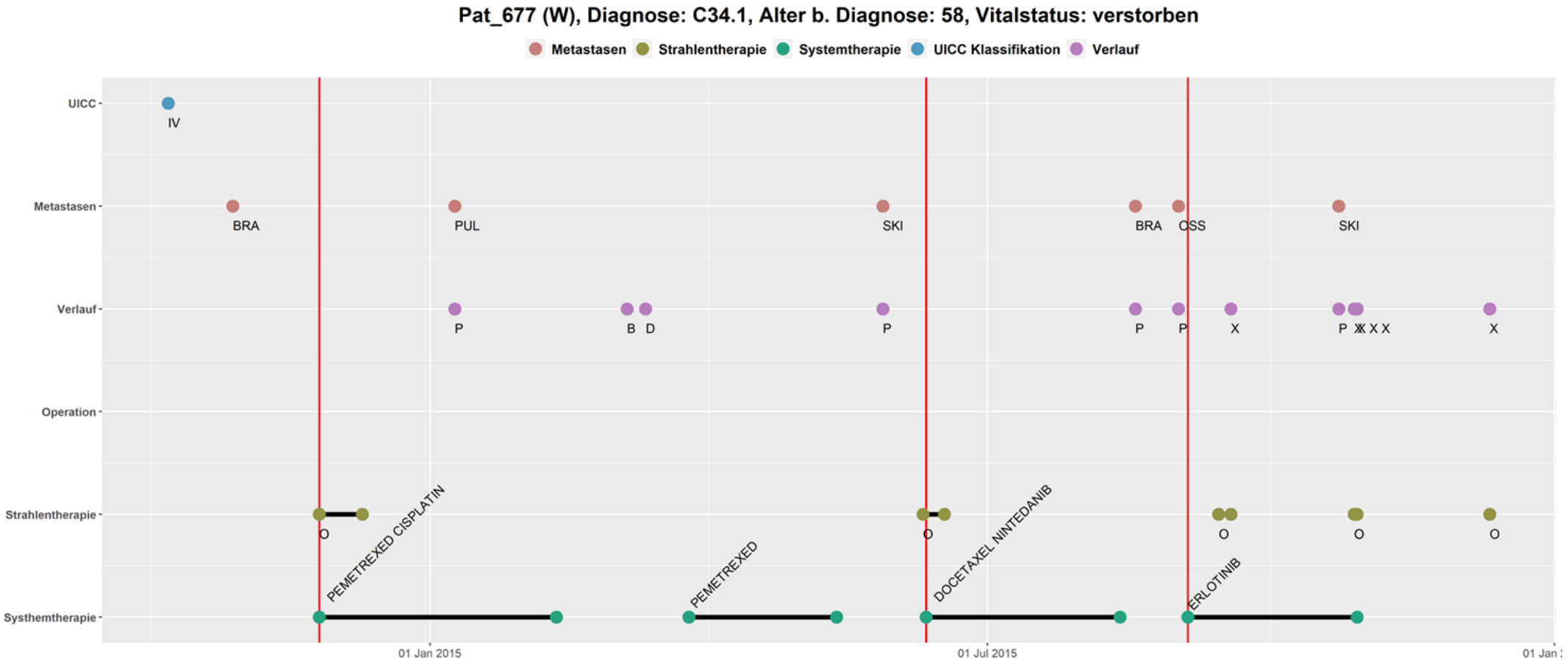
GOETHE UNIVERSITÄT FRANKFURT AM MAIN
Institut für Digitale Medizin und Klinische Datenwissenschaften




GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

# Augment data

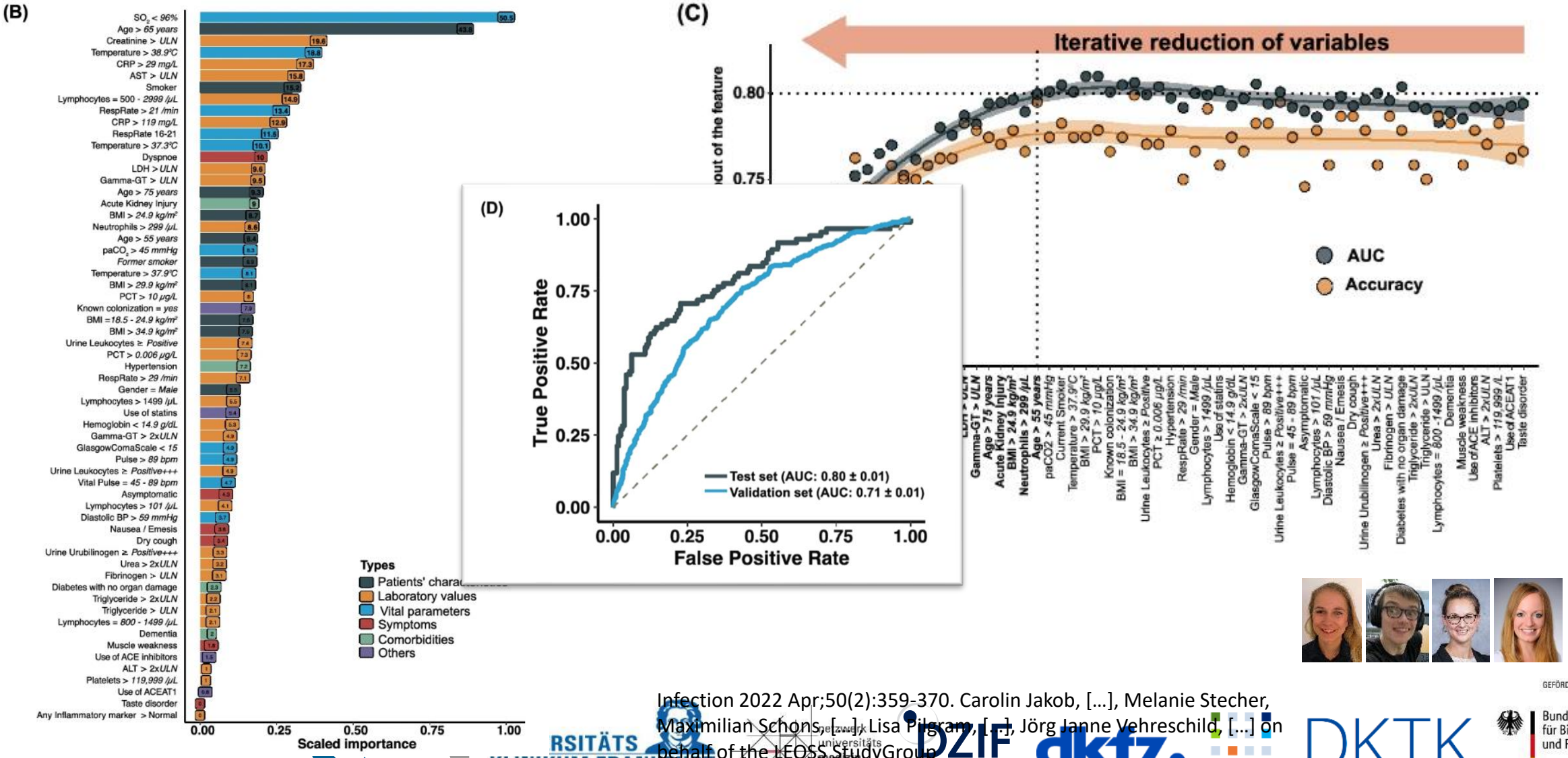
GOETHE UNIVERSITÄT FRANKFURT AM MAIN
Institut für Digitale Medizin und Klinische Datenwissenschaften




GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

# Hypothesis-free Machine Learning

GOETHE UNIVERSITÄT FRANKFURT AM MAIN
Institut für Digitale Medizin und Klinische Datenwissenschaften

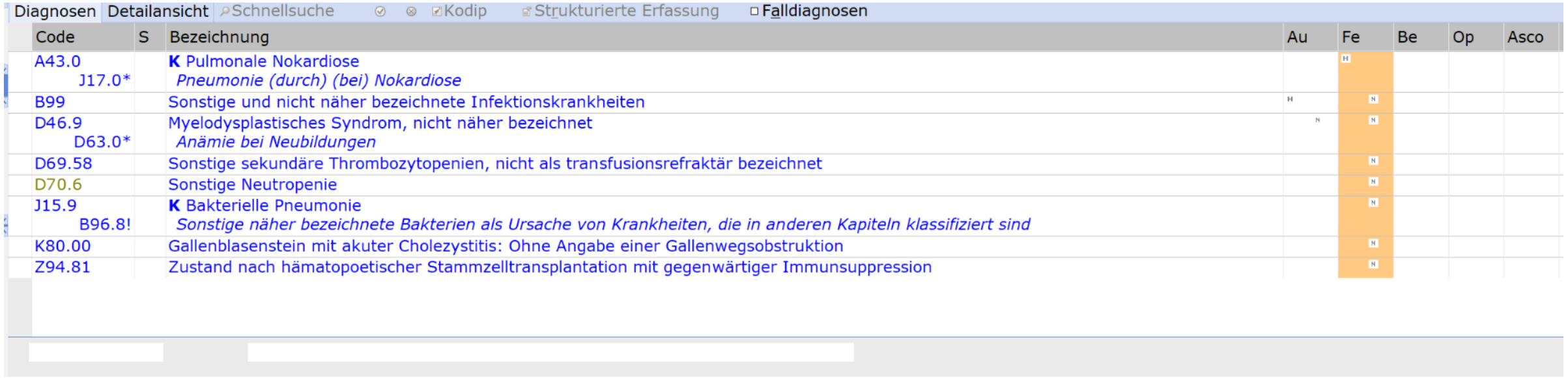





Infection 2022 Apr;50(2):359-370. Carolin Jakob, […], Melanie Stecher, Maximilian Schons, […], Lisa Pilgram, […], Jörg Janne Vehreschild, […] on behalf of the LEOSS StudyGroup

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

# 4. Limitations and Biases: Data quality



Diagnosen | Detailansicht | Schnellsuche | Kodip | Strukturierte Erfassung | Falldiagnosen

| Code | S | Bezeichnung | Au | Fe | Be | Op | Asco |
|---|---|---|---|---|---|---|---|
| A43.0<br>J17.0* | | **K** Pulmonale Nokardiose<br>*Pneumonie (durch) (bei) Nokardiose* | | H | | | |
| B99 | | Sonstige und nicht näher bezeichnete Infektionskrankheiten | H | N | | | |
| D46.9<br>D63.0* | | Myelodysplastisches Syndrom, nicht näher bezeichnet<br>*Anämie bei Neubildungen* | N | N | | | |
| D69.58 | | Sonstige sekundäre Thrombozytopenien, nicht als transfusionsrefraktär bezeichnet | | N | | | |
| D70.6 | | Sonstige Neutropenie | | N | | | |
| J15.9<br>B96.8! | | **K** Bakterielle Pneumonie<br>*Sonstige näher bezeichnete Bakterien als Ursache von Krankheiten, die in anderen Kapiteln klassifiziert sind* | | N | | | |
| K80.00 | | Gallenblasenstein mit akuter Cholezystitis: Ohne Angabe einer Gallenwegsobstruktion | | N | | | |
| Z94.81 | | Zustand nach hämatopoetischer Stammzelltransplantation mit gegenwärtiger Immunsuppression | | N | | | |

# 4. Limitations and Biases: Data quality

GOETHE UNIVERSITÄT FRANKFURT AM MAIN
Institut für Digitale Medizin und Klinische Datenwissenschaften

Diagnosen | Detailansicht | Schnellsuche | Kodip | Strukturierte Erfassung | Falldiagnosen

| Code | S | Bezeichnung |
|---|---|---|
| A43.0<br>J17.0* | | **K** Pulmonale Nokardiose<br>*Pneumonie (durch) (bei) Nokardiose* |
| B99 | | Sonstige und nicht näher bezeichnete Infektionskrankheiten |
| D46.9<br>D63.0* | | Myelodysplastisches Syndrom, nicht näher bezeichnet<br>*Anämie bei Neubildungen* |
| D69.58 | | Sonstige sekundäre Thrombozytopenien, nicht als transfusionsrefraktär bezeichnet |
| D70.6 | | Sonstige Neutropenie |
| J15.9<br>B96.8! | | **K** Bakterielle Pneumonie<br>*Sonstige näher bezeichnete Bakterien als Ursache von Krankheiten, die in anderen Kapiteln klass* |
| K80.00 | | Gallenblasenstein mit akuter Cholezystitis: Ohne Angabe einer Gallenwegsobstruktion |
| Z94.81 | | Zustand nach hämatopoetischer Stammzelltransplantation mit gegenwärtiger Immunsuppression |

**HCT-CI vor TX: 0**

**aGvHD:**
keine

**Komplikationen:**
1. Pilzpneumonie (klinische Diagnose), ausgeprägte Halluzinationen unter VFend,
2. histologisch gesichertes Basalzellkarzinom re Oberschenkel

**Transfusionsregel:**
- Erythrozyten- und Thrombozytenkonzentrate sind mit 30 Gy zu bestrahlen und CMV-frei zu transfundieren!
- EK's: 0 Rh+
- TK's: AB>B>A>0 Rh+
- FFP's: AB Rh+

**Chimärismus-Verlauf (Agendix):**
- Tag +14: 96%, 5 von 5 Empfängersignale
- Tag +30: 80%, 5 von 5 Empfängersignale

**MRD-Marker zur Verlaufskontrolle:** TET 2 (**Labor:** MLL München)

**MRD Verlauf:**
TET2 nachweisbar, U2AF1 nachweisbar (20.12.2018) ED
TET2 nicht nachweisbar, U2AF1 nachweisbar (05.03.2018) nach Ind I
TET2 nicht nachweisbar, U2AF1 nachweisbar (26.04.2018) nach allo Tx

**Leistenhernie rechts**
- aktuell (8/2018) unter intensiver immunsuppressier Therapie zunächst keine chirurgische Intervention

**Port-Implantation am 28.08.2018**

**Splenomegalie**

**Cholezystolithiasis**

**Hämorrhoiden bis IV° sowie Analprolaps**
- 2-fache Gummibandligatur bei schmerzhaften Hämorrhoiden 2. Grades am 08.08.2018
- keine erneute endoskopische Interventionsmöglichkeit am 23.08.2018 bei Hämorrhoiden IV sowie Analprolaps
- aktuell (8/2018) unter intensiver immunsuppressier Therapie zunächst keine chirurgische Intervention

**chronische Niereninsuffizienz, a.e. med.-toxischer Genese**
- Cystatin C-Clearance von 50ml/min (Befund vom 27.12.2018)

HLA-A*24:02, *26:01; HLA-B*07:02, *38:01, HLA-C07*02:, *12:03, HLA-DRB1*13:01, *16:01, HLA-DQB1*05:02,*06:03

**HLA-Retypisierung Spender (DE DKM 2963744):**
HLA-A*24:02, *26:01; HLA-B*07:02, *38:01, HLA-C07*02:, *12:03, HLA-DRB1*13:01, *16:01, HLA-DQB1*05:02,*06:03

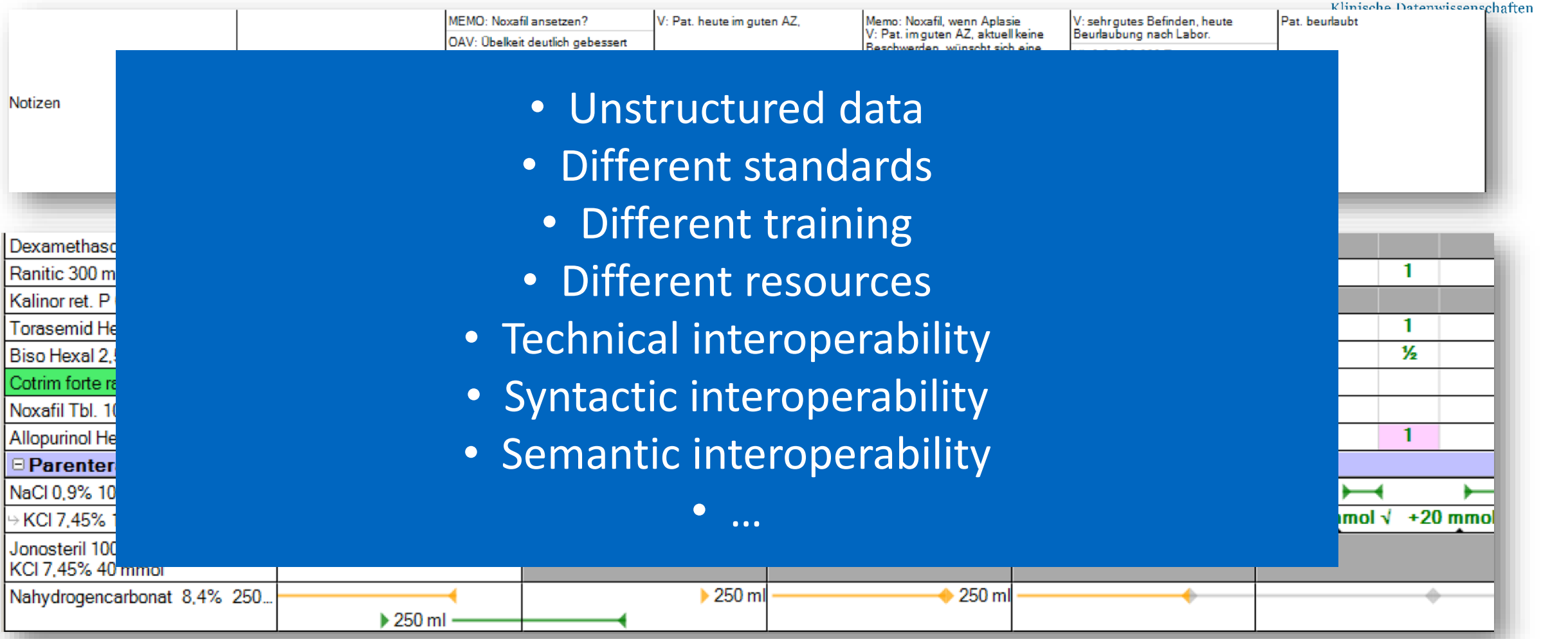
**HLA-Antikörper (Luminex):** nicht erforderlich

**Remissionsstatus vor TX (KMP vom 5.03.2018):** Histopathologie: Eine reifungsgestörte Hämatopoese mit Stromaödem, entzündlicher Markraumreaktion und Persistenz einer CD34- positiven Progenitorzellpopulation von knapp über 5% der kernhaltigen Zellen neben einer initialen Vermehrung retikulärer Knochenmarksfasern (fokal MF-1).

**MRD: MLL-München: keine Nachweis TET2, U2AF1 persistierend nachweisbar**

**HCT-CI vor TX: 0**

# 4. Limitations and Biases: Data quality




- Unstructured data
- Different standards
- Different training
- Different resources
- Technical interoperability
- Syntactic interoperability
- Semantic interoperability
- ...

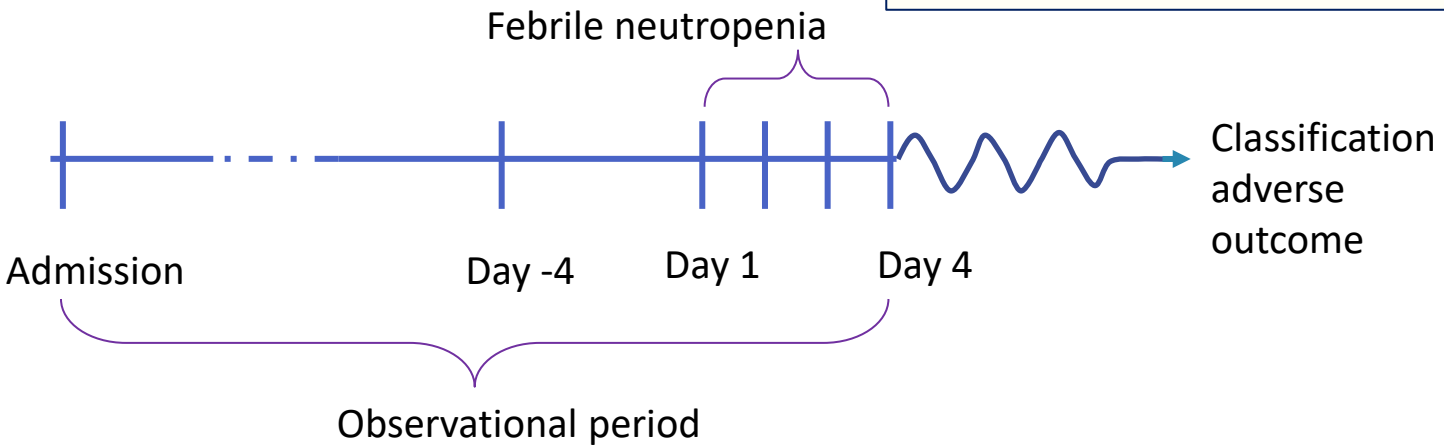
# Example: Machine Learning in Neutropenic Fever

GOETHE UNIVERSITÄT FRANKFURT AM MAIN
Institut für Digitale Medizin und Klinische Datenwissenschaften

Laboratory data (145)
Body temperature (19)
Microbiological findings (23)
Virological findings (8)
Radiological findings (3)
Patient characteristics (8)
Underlying disease
Neutropenia (6)
Medication (9)
Comorbidities (4)

→ Value for day 1 to day 4 (febrile neutropenia (FN))
→ Value for day -4 to day 0 (time before FN)
→ Minimum/maximum value
→ Days from minimum/maximum value to day 1
→ Differences of values

→ Choice of class of features/categorisation
→ Missing data handling/imputation
→ Visualise interaction of features



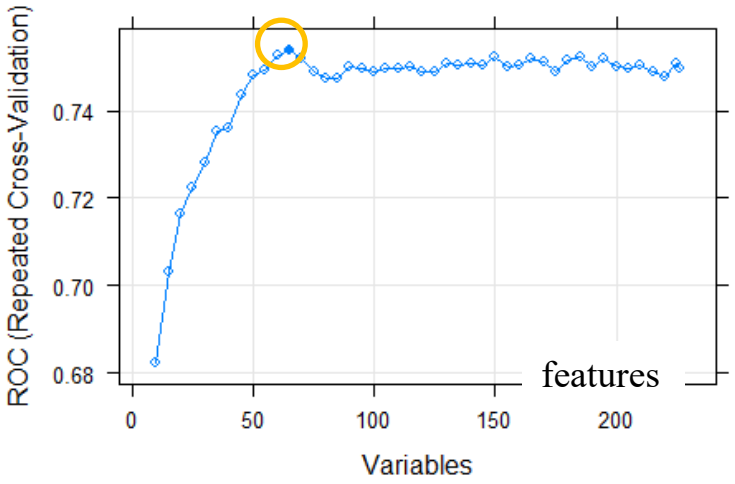

Jakob C, Vehreschild J, ECCMID 2019

GEFÖRDERT VOM Bundesministerium für Bildung und Forschung

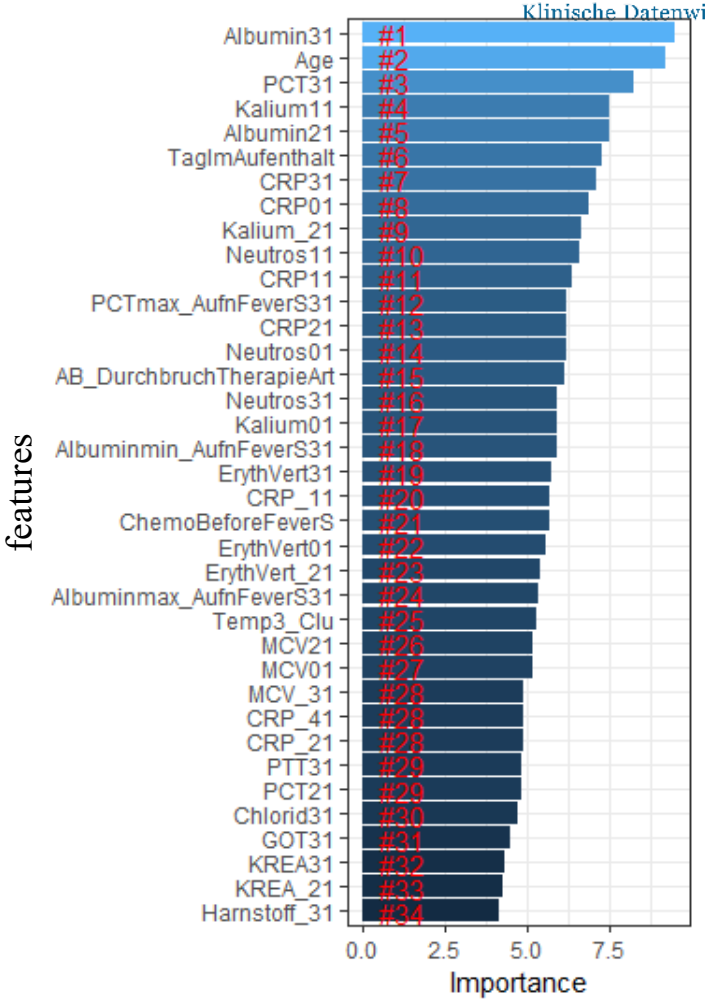
# Example: Machine Learning in Neutropenic Fever

GOETHE UNIVERSITÄT FRANKFURT AM MAIN
Institut für Digitale Medizin und Klinische Datenwissenschaften

- 65 selected features




| predictions | actual YES | NO |
|---|---|---|
| YES | 8 | 1 |
| NO | 14 | 101 |




- Internal validation AUC = 0.75
- Out-of-sample validation AUC = 0.68

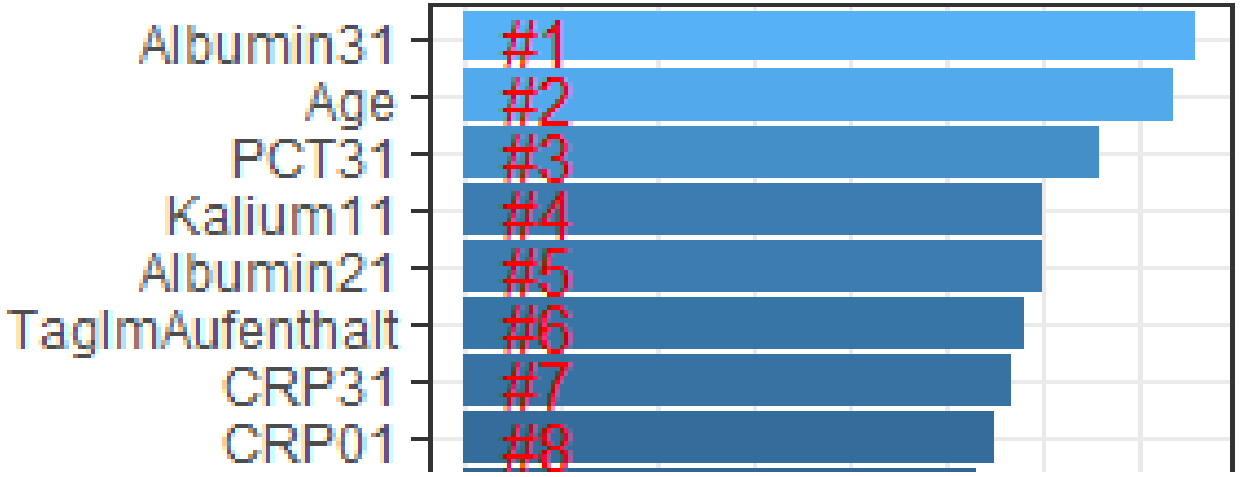
Jakob C, Vehreschild J, ECCMID 2019

GEFÖRDERT VOM Bundesministerium für Bildung und Forschung

# Example: Machine Learning in Neutropenic Fever

GOETHE UNIVERSITÄT FRANKFURT AM MAIN
Institut für Digitale Medizin und Klinische Datenwissenschaften

- 65 selected features




- Internal validation AUC = 0.75
- Out-of-sample validation AUC = 0.68

Jakob C, Vehreschild J, ECCMID 2019

GEFÖRDERT VOM Bundesministerium für Bildung und Forschung

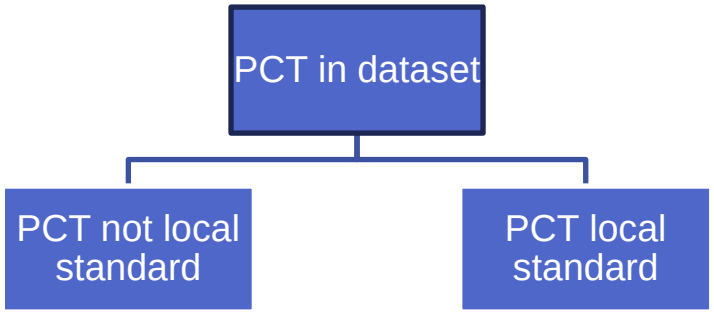
# Possible Meanings of Missingness

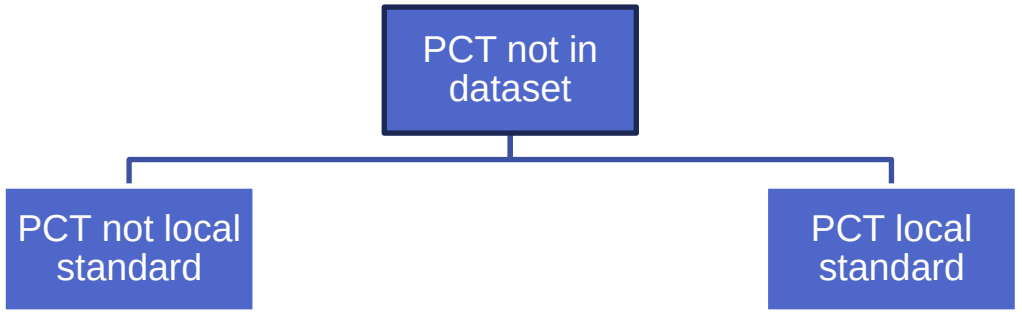
- Performed in another center / department / location / data system
- Data loss
- Unstructured / cryptic documentation
- Hand-written note
- Lack of interface / data transfer process
- Intentionally not done (not needed, too expensive / lack of reimbursement or result obvious)
- Unintentionally not done (forgotten, sample lost, unable to perform procedure)
- Done, but intentionally not documented (forensic issues)
- Done, but unintentionally not documented (failed to record / trsnscribe)

# Examples for Possible Interpretations of Missingness

GOETHE UNIVERSITÄT FRANKFURT AM MAIN
Institut für Digitale Medizin und Klinische Datenwissenschaften

- A patient with community-acquired pneumonia presents himself in the emergency department
- The inflammatory paremeter „Procalcitonin“ offers good risk classification, but is expensive




Indicator: Quality of care

Indicator: Clinical condition

Random errors

Direct Interpretation Possible

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

# Missingness (and presence) of Data in Real-World Setting

GOETHE UNIVERSITÄT FRANKFURT AM MAIN
Institut für Digitale Medizin und Klinische Datenwissenschaften

**Missing completely at random (MCAR)**

**Missing at random (MAR)**

**Not missing at random (NMAR)**

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

# Missing completely at random (MCAR)

GOETHE UNIVERSITÄT FRANKFURT AM MAIN
Institut für Digitale Medizin und Klinische Datenwissenschaften

**„Data is missing for no obvious reason“**

- Does not introduce bias
- Mass MCAR missingness may cause loss of power
- Mass MCAR may cause underestimation of effect sizes

Faulty data interface at one of multiple study sites

Random documentation mistakes

Random treatment mistakes (missing prescription, diagnostic test)

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

# Missing at random (MAR)

GOETHE UNIVERSITÄT FRANKFURT AM MAIN
Institut für Digitale Medizin und Klinische Datenwissenschaften

**„Missingness is related to a variable outside the primary observation“**

- May cause biased overall results
- Relationship between variables intact

Smaller hospitals less likely to order expensive tests/drugs

Concurring study leads to documentation focus on distinct population

Less comprehensive documentation in elderly / terminally ill patients

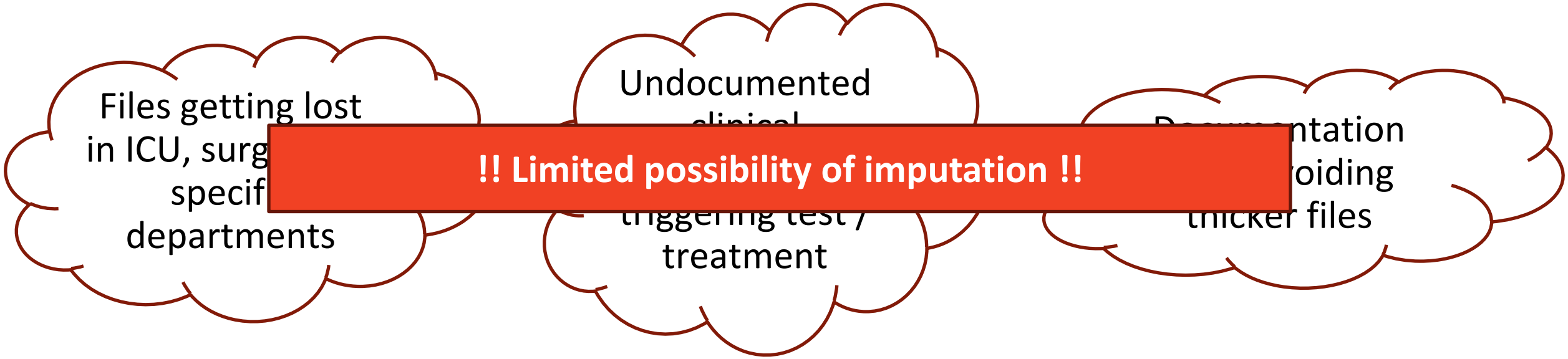
GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

# Not missing at random (NMAR)

GOETHE UNIVERSITÄT FRANKFURT AM MAIN
Institut für Digitale Medizin und Klinische Datenwissenschaften

**„Missingness is related to the primary observation“**

- Causes biased overall results
- Causes biased relationship between variables




**!! Limited possibility of imputation !!**

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

# Handling Missingness

- **„Complete record analyses“** = Drop everything with one missing variable
  - May cause bias in NMAR scenarios
  - Greatest loss of power
- **Create dummy variable / feature for missingness**
  - Causes co-linearity between dummy variable/feature and value
  - Great loss of power in MCAR scenarios
  - Good solution for sensitivity analyses
- **Impute missing values**
  - By definition limited to MAR and MCAR scenarios
  - May increase pre-existing bias in the dataset

GEFÖRDERT VOM

# Means of imputation

GOETHE UNIVERSITÄT FRANKFURT AM MAIN
Institut für Digitale Medizin und Klinische Datenwissenschaften

| Age | Treatment Group | Stage | Response | TTP (days) | Survival (days) |
|---|---|---|---|---|---|
| 56 | A | IIIb | PD | 117 | 180 |
| 62 | B | IV | SD | 100 | 210 |
| 47 | A | IV | PR | 150 | 320 |
| 65 | B | IIIb | CR | 180 | 400 |
| 59 | A | IV | PD | 117 | 150 |
| 53 | B | IIIb | PR | 200 | 365 |
| 61 | A | IV | SD | 90 | 200 |
| 58 | B | IV | PD | 80 | 160 |
| 49 | A | IIIb | PR | 120 | 300 |
| 67 | B | IV | CR | 250 | 500 |

- Mean Value

Pseudo-exactness, artificially narrow CIs, biased histogramm, loss of effect size, masqued interactions, unrealistic values…

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

# Means of imputation

| Age | Treatment Group | Stage | Response | TTP (days) | Survival (days) |
|---|---|---|---|---|---|
| 56 | A | IIIb | PD | 110 | 180 |
| 62 | B | IV | SD | 100 | 210 |
| 47 | A | IV | PR | 150 | 320 |
| 65 | B | IIIb | CR | 180 | 400 |
| 59 | A | IV | PD | 110 | 150 |
| 53 | B | IIIb | PR | 200 | 365 |
| 61 | A | IV | SD | 90 | 200 |
| 58 | B | IV | PD | 80 | 160 |
| 49 | A | IIIb | PR | 120 | 300 |
| 67 | B | IV | CR | 250 | 500 |

- Mean Value
- Class/group based mean

Pseudo-exactness, artificially narrow CIs, biased histogramm , unrealistic values …

# Means of imputation

GOETHE UNIVERSITÄT FRANKFURT AM MAIN
Institut für Digitale Medizin und Klinische Datenwissenschaften

| Age | Treatment Group | Stage | Response | TTP (days) | Survival (days) |
|---|---|---|---|---|---|
| 56 | A | IIIb | PD | 83 | 180 |
| 62 | B | IV | SD | 100 | 210 |
| 47 | A | IV | PR | 150 | 320 |
| 65 | B | IIIb | CR | 180 | 400 |
| 59 | A | IV | PD | 54 | 150 |
| 53 | B | IIIb | PR | 200 | 365 |
| 61 | A | IV | SD | 90 | 200 |
| 58 | B | IV | PD | 80 | 160 |
| 49 | A | IIIb | PR | 120 | 300 |
| 67 | B | IV | CR | 250 | 500 |

- Mean Value
- Class/group based mean
- Model-based

Pseudo-exactness, artificially narrow CIs

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

# Means of imputation

| Age | Treatment Group | Stage | Response | TTP (days) | Survival (days) |
|---|---|---|---|---|---|
| 56 | A | IIIb | PD | | 180 |
| 56 | A | IIIb | PD | 56 | 180 |
| 56 | A | IIIb | PD | 67 | 180 |
| 56 | A | IIIb | PD | 78 | 180 |
| 56 | A | IIIb | PD | 93 | 180 |
| 56 | A | IIIb | PD | 45 | 180 |
| 56 | A | IIIb | PD | 67 | 180 |
| 56 | A | IIIb | PD | 110 | 180 |
| 56 | A | IIIb | PD | 78 | 180 |
| 56 | A | IIIb | PD | 64 | 180 |
| 56 | A | IIIb | PD | 98 | 180 |
| 56 | A | IIIb | PD | 78 | 180 |

- Mean Value
- Class/group based mean
- Model-based
- Multiple imputations +/- chained equations

High workload, possible bias by regression models, maintains bias in MNAR scenarios

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

# Causal Machine Learning?

GOETHE UNIVERSITÄT FRANKFURT AM MAIN
Institut für Digitale Medizin und Klinische Datenwissenschaften

nature medicine

Perspective

https://doi.org/10.1038/s41591-024-02902-1

## Causal machine learning for predicting treatment outcomes

Received: 3 January 2024

Accepted: 4 March 2024

Stefan Feuerriegel[1,2] ✉, Dennis Frauen[1,2], Valentyn Melnychuk[1,2], Jonas Schweisthal[1,2], Konstantin Hess[1,2], Alicia Curth[3], Stefan Bauer[4,5], Niki Kilbertus[2,4,5], Isaac S. Kohane[6] & Mihaela van der Schaar[7,8]

Feuerriegel S et al.,
Nature Medicine 2024,
doi: 10.1038/s41591-024-02902-1

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

# Working with Real-World Data

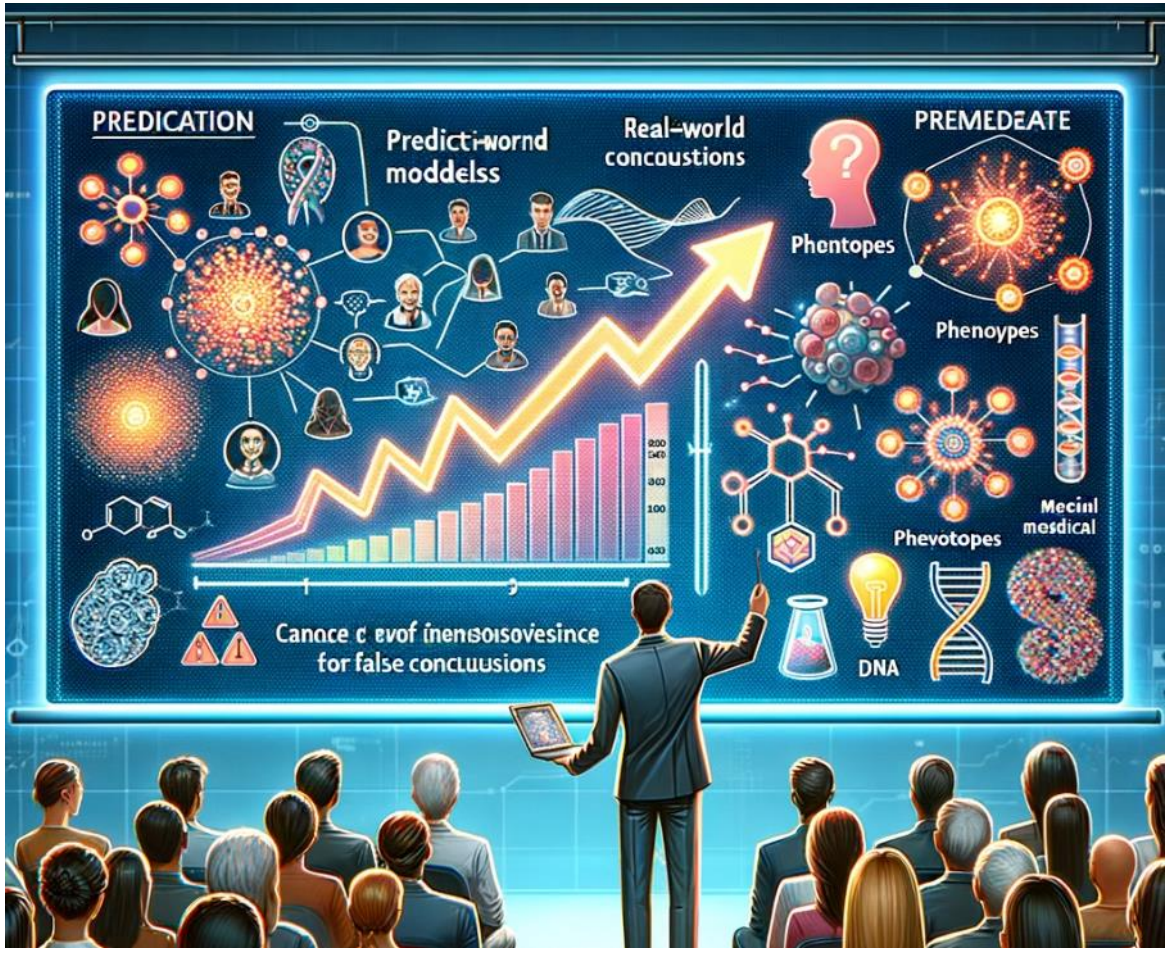
GOETHE UNIVERSITÄT FRANKFURT AM MAIN
Institut für Digitale Medizin und Klinische Datenwissenschaften

- Consult a clinician
- Understand your data and where it comes from
- Prepare and compare your data
- Perform EXTENSIVE sensitivity analyses
- For prediction, only use features with high availability, normalize timelines

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

# Take Home Message

- We are about(ish) to enter a new age of clinical data availability
- Real-world data = abundant & powerful
- Real-word data also = laborious & difficult to process
- Risk of false conclusions (prediction models!)
- Chance of new discoveries (phenotypes! precision medicine!)




GEFÖRDERT VOM