

Session II: Praktische Arbeit mit tranSMART

Aufbereitung und Analyse von Daten in phänotypischen
und genotypischen Forschungsdatenbanken mit
tranSMART | 05.08.2016 | Berlin



Teil 1

Ein kurzes einführendes Szenario mit tranSMART

Aufgabe 1.1

Als motivierendes Beispiel soll ein Signifikanztest (exakter Fisher-Yates-Test oder exakter Chi-Quadrat-Test) an einem Datensatz aus einer Brustkrebsstudie durchgeführt werden. Ein Fishertest liefert ein Maß für die Unabhängigkeit zweier Merkmale.

- ▶ Wie unabhängig ist die Überlebenszeit von der Tumorgroße?

Anmelden am System



Please login...

Username :

Password :

Remember me

:

Not a user ? Contact [administrator](#) to request an account

- Instanz 1:
 - URL: <http://postgres-demo.transmartfoundation.org/transmart>
 - uid: admin
 - pwd: admin
- Instanz 2:
 - URL: <http://public.etriks.org>
 - uid: guest
 - pwd: transmart2015

Browse: Studienübersicht und Metadaten



Navigation: All | Export Cart | **Browse** | Analyze | Sample Explorer | Gene Signature/Lists | GWAS | Admin | Utilities

Active Filters: and | Filter | Clear

Program Explorer

- function_test_progroma
- MAGIC_Manning_et_al
- Public Studies
 - Asthma_Choy_GSE23611
 - Asthma_Tsitsiou_GSE31773
 - Brain_Cancer_Phillips_GSE4271
 - Breast_Cancer_Ishvina_GSE4922
 - Breast_Cancer_Minn_GSE5327
 - Breast_Cancer_Pawitan_GSE1456
 - Breast_Cancer_Sorlie_GSE4382
 - COPD_Bhattacharya_GSE8581
 - Diabetes_Type_2_Taneera_GSE38642
 - GSE13168 Effects of glucocorticoids and
 - GSE3446 Airway Epithelial miRNA Expre
 - GSE4698 Molecular characterization of v
 - GSE48213
 - Melanoma_Uveal_Gangemi_GSE27831
 - Psoriasis_Zaba_GSE11903
 - Pulmonary Sarcoidosis_Ho_GSE19976
 - Rheumatoid_Arthritis_Andreas_GSE1002
 - Rheumatoid_Arthritis_Takeuchi_GSE206
 - Rheumatoid_Arthritis_Yarilina_GSE10500
 - TEST_GSE48213_1
 - test_RNAseq_GSE48213
- Test Studies

Study: **Breast_Cancer_Sorlie_GSE4382**

Buttons: Add new analysis | Add new assay | Add new folder

Repeated observation of breast tumor subtypes in independent gene expression data sets. Characteristic patterns of gene expression measured by DNA microarrays have been used to classify tumors into clinically relevant subgroups. In this study, we have refined the previously defined subtypes of breast tumors that could be distinguished by their distinct patterns of gene expression. A total of 115 malignant breast tumors were analyzed by hierarchical clustering based on patterns of expression of 534 "intrinsic" genes and shown to subdivide into one basal-like, one ERBB2-overexpressing, two luminal-like, and one normal breast tissue-like subgroup. The genes used for classification were selected based on their similar expression levels between pairs of consecutive samples taken from the same tumor separated by 15 weeks of neoadjuvant treatment. Similar cluster analyses of two published, independent data sets representing different patient cohorts from different laboratories, uncovered some of the same breast cancer subtypes. In the one data set that included information on time to development of distant metastasis, subtypes were associated with significant differences in this clinical feature. By including a group of tumors from BRCA1 carriers in the analysis, we found that this genotype predisposes to the basal tumor subtype. Our results strongly support the idea that many of these breast tumor subtypes represent biologically distinct disease entities. This SuperSeries is composed of the SubSeries: GSE4335 Norway/Stanford Breast Tumors GSE4336 Breast tumors

Subject-level data is available for this study. **Open in Analyze view**

Associated Tags

Property	Value
Study identifier	GSE4382
Pathology	Breast Neoplasms
Study phase	Not applicable
Study objective	Not applicable
Study design	Other
Study biomarker type	Not applicable
Study link	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE4382
Number of followed subjects	167
Organism	Homo sapiens
Study access type	Public
Country	UNITED STATES
Study date	Mar 07, 2006
Study PubMed ID	12829800
Study publication DOI	10.1073/pnas.0932692100
Study publication author list	Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, Demeter J, Perou CM, Lonning PE, Brown PO, Borresen-Dale AL, Botstein D.
Study publication title	Repeated observation of breast tumor subtypes in independent gene expression data sets.
Study publication status	Published

Report a problem

Links: Navigationsbaum für klinische Konzepte



The screenshot shows a web-based interface for navigating clinical concepts. On the left is a 'Navigate Terms' tree view for 'Breast Cancer_Sorlie_GSE4382 (167)'. The tree is organized into folders (Ordner) such as 'Public Studies', 'Biomarker Data', 'Gene Expression (167)', 'Subjects', 'Demographics (118)', 'End Points (118)', 'Medical History (167)', 'Tumor Characteristics (167)', and 'Vital Status (167)'. Red dashed arrows point to specific levels: 'Ordner' points to the tree structure, 'Hochdim.' points to 'Gene Expression (167)', 'Metrisch' points to 'Age (118)', and 'Kategorial' points to 'Vital Status (167)'. The right side of the interface is a 'Subset' editor with two columns, 'Subset 1' and 'Subset 2'. Each column contains a list of selected terms with 'Exclude' and 'X' buttons. The top navigation bar includes 'Browse', 'Analyze', 'Sample Explorer', 'Gene Signature/Lists', 'GWAS', 'Admin', and 'Utilities'. A 'Save Subset' button is visible in the top right.

- Ordner dienen zur Strukturierung (aufklappbar, enthalten Unterordner und/oder Merkmale)
- Merkmale können kategorial, stetig/metrisch oder hochdimensional sein
- Anordnung der Merkmale ist wahlfrei (Spezifikation bei Import), aber für Verständnis Dritter wichtig
- Konzepte werden angeklickt und in das Subset-Fenster gezogen (im Falle eines Ordner umfasst dies alle enthaltenen Konzepte)

Report a problem

Auswahl der gesamten Studie

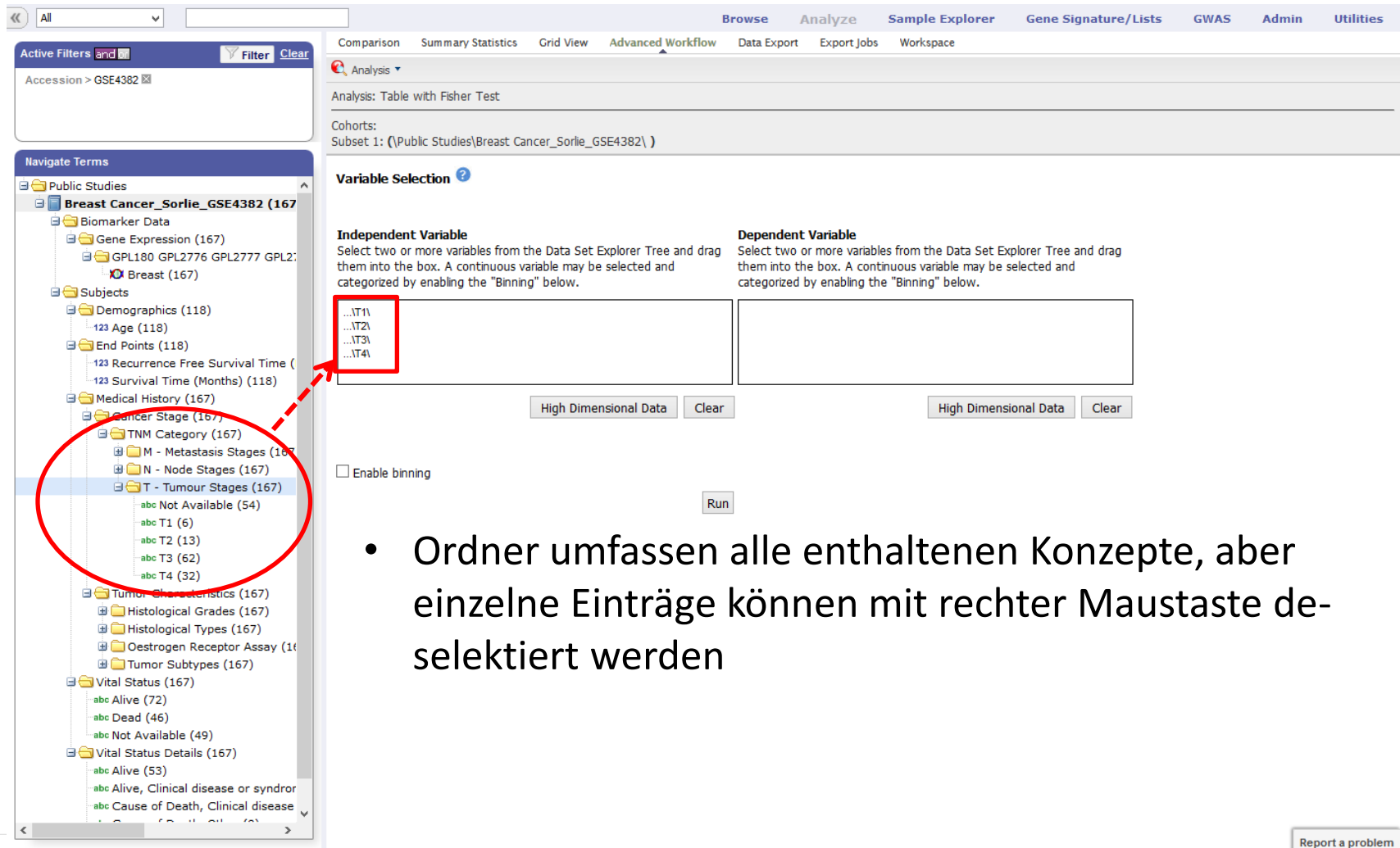


The screenshot shows the TMF interface with the following elements:

- Active Filters:** Accession > GSE4382
- Navigate Terms:** A tree view showing the hierarchy of terms. The term **Breast Cancer_Sorlie_GSE4382 (167)** is highlighted with a red oval.
- Subset 1:** A list of selected terms. The term **Breast Cancer_Sorlie_GSE4382** is highlighted with a red box. Below it are AND connectors and Exclude/X buttons.
- Subset 2:** An empty list for a second subset.
- Buttons:** Save Subset, Clear, Filter, and Clear buttons are visible.

- Die gewählte Kohorte umfasst alle Probanden, genauer gesagt alle Fakten, die in der DB zu dieser Studie abgespeichert sind
- Weitere Konzepte können in das gleiche Fenster (ODER-Verknüpfung) in das darunter (UND-Verknüpfung) oder das daneben (2. Kohorte für Vergleiche) geschoben werden
- Näheres in den praktischen Übungen

Es müssen die zwei zu untersuchenden Konzepte ausgewählt werden: 1. Tumorgröße nach TNM



Active Filters and

Accession > GSE4382

Navigate Terms

- Public Studies
 - Breast Cancer_Sorlie_GSE4382 (167)
 - Biomarker Data
 - Gene Expression (167)
 - GPL180 GPL2776 GPL2777 GPL2778
 - Breast (167)
 - Subjects
 - Demographics (118)
 - Age (118)
 - End Points (118)
 - Recurrence Free Survival Time (118)
 - Survival Time (Months) (118)
 - Medical History (167)
 - Cancer Stage (167)
 - TNM Category (167)
 - M - Metastasis Stages (167)
 - N - Node Stages (167)
 - T - Tumour Stages (167)**
 - Not Available (54)
 - T1 (6)
 - T2 (13)
 - T3 (62)
 - T4 (32)

Comparison Summary Statistics Grid View **Advanced Workflow** Data Export Export Jobs Workspace

Analysis

Analysis: Table with Fisher Test

Cohorts:

Subset 1: (Public Studies\Breast Cancer_Sorlie_GSE4382)

Variable Selection

Independent Variable
Select two or more variables from the Data Set Explorer Tree and drag them into the box. A continuous variable may be selected and categorized by enabling the "Binning" below.

Dependent Variable
Select two or more variables from the Data Set Explorer Tree and drag them into the box. A continuous variable may be selected and categorized by enabling the "Binning" below.

...T1
...T2
...T3
...T4

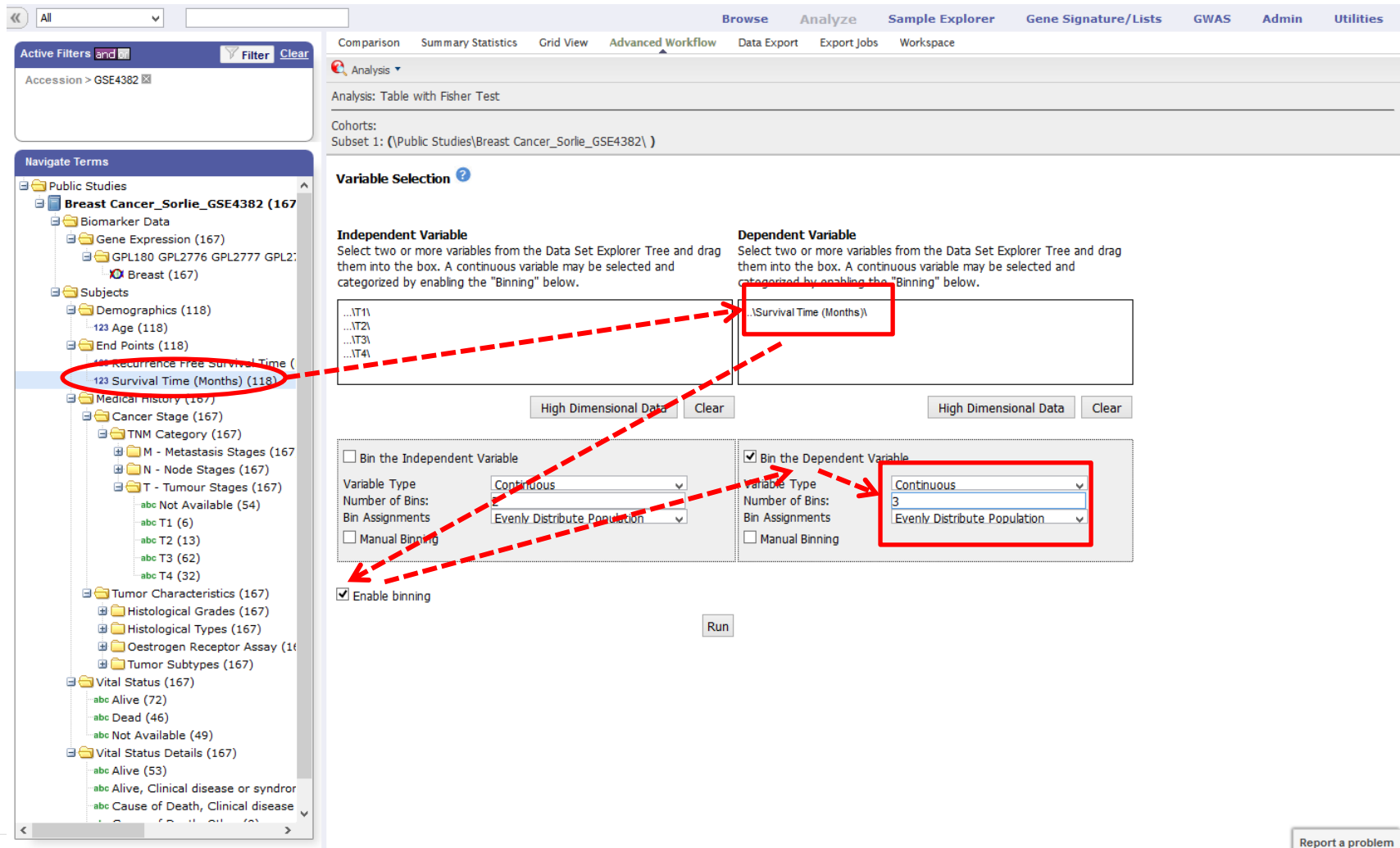
High Dimensional Data Clear High Dimensional Data Clear

Enable binning

Run

- Ordner umfassen alle enthaltenen Konzepte, aber einzelne Einträge können mit rechter Maustaste de-selektiert werden

2. Überlebenszeit: das ist kein kategoriales Merkmal, aber tranSMART kann es dazu machen!



The screenshot shows the tranSMART interface with the following elements:

- Active Filters:** Accession > GSE4382
- Navigate Terms:** A tree view on the left with 'Survival Time (Months)' (118) circled in red.
- Variable Selection:**
 - Independent Variable:** A box containing 'Survival Time (Months)' (highlighted with a red box).
 - Dependent Variable:** A box containing 'Survival Time (Months)' (highlighted with a red box).
 - Bin the Independent Variable:**
 - Variable Type: Continuous
 - Number of Bins: 2
 - Bin Assignments: Evenly Distribute Population
 - Bin the Dependent Variable:**
 - Bin the Dependent Variable
 - Variable Type: Continuous
 - Number of Bins: 3
 - Bin Assignments: Evenly Distribute Population
 - Enable binning
 - Run** button

Report a problem

Ladies and Gentlemen, please start your engines!



The screenshot displays the transSMART software interface. On the left, the 'Navigate Terms' tree shows a hierarchy for 'Breast Cancer_Sorlie_GSE4382 (167)', with 'Survival Time (Months) (118)' selected. The top navigation bar includes 'Browse', 'Analyze', 'Sample Explorer', 'Gene Signature/Lists', 'GWAS', 'Admin', and 'Utilities'. The main panel shows 'Analysis: Table with Fisher Test' and 'Subset 1: (\Public Studies\Breast Cancer_Sorlie_GSE4382\)'. The 'Variable Selection' section is active, with 'Independent Variable' and 'Dependent Variable' fields. The 'Independent Variable' field contains '...T1\', '...T2\', '...T3\', and '...T4\'. The 'Dependent Variable' field contains '...\Survival Time (Months)\'. Below these fields are 'High Dimensional Data' and 'Clear' buttons. The 'Bin the Independent Variable' section has 'Bin the Independent Variable' unchecked, 'Variable Type' set to 'Continuous', 'Number of Bins' set to '2', and 'Bin Assignments' set to 'Evenly Distribute Population'. The 'Bin the Dependent Variable' section has 'Bin the Dependent Variable' checked, 'Variable Type' set to 'Continuous', 'Number of Bins' set to '3', and 'Bin Assignments' set to 'Evenly Distribute Population'. A red dashed arrow points to the 'Run' button, which is highlighted with a red box.

Report a problem

Ergebnis: p-Wert nach Fischer 0,278; p-Wert nach χ^2 -Test 0,272

The screenshot shows a software interface with a left-hand navigation tree and a main analysis panel. The navigation tree includes categories like 'Public Studies', 'Breast Cancer', and 'Survival Time (Months)'. The main panel shows analysis settings for binning variables. Below the settings is a contingency table and a summary of statistical tests.

Gebildete Altersgruppen

	17.00000 < Y ≤ 39.00000	3.00000 ≤ Y ≤ 17.00000	39.00000 < Y ≤ 188.00000
T1	2	3	1
T2	2	7	4
T3	19	18	25
T4	14	11	7

Kontingenztafel

Fisher test p-value	0.278
χ^2	7.56
χ^2 p-value	0.272

[Download raw R data](#)

- Interpretation über Signifikanzniveau, d.h. Nullhypothese der Unabhängigkeit kann nicht verworfen werden für ein Signifikanzniveau $\alpha = 0,05$

Webinars und Tutorials: umfangreiche Ressourcen verfügbar!



The screenshot shows the tranSMART Foundation website. At the top left is the logo for the tranSMART FOUNDATION, which includes a DNA double helix icon. To the right of the logo are social media icons for Twitter, LinkedIn, Facebook, Google+, and YouTube. Below these are 'Home' and 'Contact Us' buttons. A dark blue navigation bar contains the following menu items: ABOUT US, NEWS, EVENTS, MEMBERSHIP, RESEARCHERS, PLATFORM, MARKETPLACE, and BLOG. A dropdown menu is open under 'RESEARCHERS', listing: OVERVIEW OF RESEARCH USES, TRAINING & TUTORIALS (highlighted with a blue arrow), TRANSMART BIBLIOGRAPHY, CURATED DATASETS, and USE CASES ON WIKI. A second dropdown menu is open under 'TRAINING & TUTORIALS', listing: THE 2016 TRANSMART FOUNDATION TRAINING PROGRAM, RECORDINGS OF TRAINING CLASSES (highlighted with a red box), and TUTORIALS. A red dashed arrow points from the top right towards the 'RECORDINGS OF TRAINING CLASSES' option. The main content area features a section titled 'tranSMART Foundation Training & Tutorial' with a sub-header 'LAST UPDATED 11 Feb 2016: Link to 2016 Training Program; added'. Below this is a paragraph: 'The tranSMART Foundation and members have created a training & tutorial program on the tranSMART Platform to answer clinical genomic questions. We are growing a set of collateral materials in the form of documents, training manuals and video tutorials that are available on this site. In addition, a regular program of training offerings is being created as well as the opportunity to request special or customized training for your organization or project.' Further down, there is a section titled 'Training Classes' with the text: 'The 2016 tranSMART Foundation Training Program can be found on our [Training Page](#). Recordings and slide decks from previous training webinars can be found [HERE](#). Having a Training need, question or suggestion? Let us know by contacting us at [CLICK HERE](#).' At the bottom, there is a section titled 'Video Tutorials Available on YouTube' with a link to transmartfoundation.org/transmart-training-program/.

transmartfoundation.org/transmart-training-program/

Vielen Dank!

Göttingen

Prof. Dr. Ulrich Sax

Ulrich.Sax@med.uni-goettingen.de

Christian Bauer

christian.bauer@med.uni-goettingen.de

Benjamin Baum

benjamin.baum@med.uni-goettingen.de

Institut für Medizinische
Informatik,
Universitätsmedizin
Göttingen

Erlangen

Dr. Thomas Ganslandt

Thomas.Ganslandt@uk-erlangen.de

Christian Knell

christian.knell@fau.de

Jan Christoph

jan.christoph@fau.de

Lehrstuhl für Medizinische
Informatik, Friedrich-
Alexander-Universität
Erlangen-Nürnberg

Leipzig

Matthias Löbe

matthias.loebe@imise.uni-leipzig.de

Sebastian Stäubert

sebastian.staebert@imise.uni-leipzig.de

Institut für Medizinische
Informatik, Statistik und
Epidemiologie (IMISE),
Universität Leipzig