



St. Anna Kinderkrebsforschung
CHILDREN'S CANCER RESEARCH INSTITUTE



Data scraping, database implementation and visualization for analyzability of flow cytometry data related to pediatric stem cell transplantation

With emphasis on visualization and retrospective, explorative
analysis of naïve CD4+ T-cells

Master Thesis

For attainment of the academic degree of
Master of Science in Engineering (MSc)

in the Master Program Digital Healthcare
at St. Pölten University of Applied Sciences

by

Jakob Winkler, BSc

dh151824, 1510756824

First advisor: Dipl.-Sporting. Dr. Mario Heller

Second advisor: Priv.- Doz. Mag. Dr. René Geyeregger

[Vienna, 21.07.2017]

Preface

Since 04/2010, I am dedicating my profession as a pediatric nurse with additional technical background, to children and adolescences, suffering from severe diseases of the blood building system. At the stem cell transplantation ward of the St. Anna Children's hospital of Vienna children and adolescences undergo stem cell transplantations for curing these diseases. Beside empathy toward the individual patient, continuous monitoring of different blood parameters is important during the whole treatment. Reports of cell analysis are transmitted to the ward by fax, which does not conform to state of the art-technology, in my opinion.

Therefore I presented my new acquired knowledge of data processing and visualization techniques, to improve the representation of these reports, to scientist René Geyeregger, PhD. So, during an internship from December 2016 to August 2017 at the St. Anna Children Cancer Research Institute this master thesis and the related work of data transfer out of historical laboratory results into a relational database, web-based visualization technique and analysis were performed. Thank you, René, for this opportunity and your personal support!

Acknowledgements to the entire team of the laboratory for Clinical Cell Biology & FACS Core Unit, especially to Gerhard Fritsch, PhD, René Geyeregger, PhD, Dijana Trbojevic and Dieter Printz who had helped me during the whole work related to data quality assurance proofing. Thank you to Herbert Pichler, MD and Susanne Matthes-Leodolter, MD, of the St. Anna's Children Hospital, who had motivated me to develop this work and had always an open ear for my questions and issues.

The explorative statistic was a challenging part for me. Evgenia Glogova, MSc and Mario Heller, PhD, with your weekly dose of inspiration and guidance you were a great help to me!

Thank you to Mayur Parihar, MD for his hours of proof reading!

**Dedicated to my dear wife Katharina
and my lovely daughters Johanna and Veronika.
Thank you for your patience!**

Declaration

I declare that I have developed and written the enclosed Master Thesis completely by myself, and have not used sources or means without declaration in the text. Any thoughts from others or literal quotations are clearly marked. This work was not used in the same or in a similar version to achieve an academic grading or is being published elsewhere.

Vienna, 21/07/2017

Place, Date



Signature

Abstract

Flow cytometry is a commonly used technique to analyze different cell types of human blood cells by their specific surface characteristics. Furthermore, it is used to monitor cell development (engraftment) after allogenic stem cell transplantation (allo-HSCT). Since 1995, these analyses have been performed at the St. Anna Children's Cancer Research, but data has been stored with undefined data structure in multiple Excel tables. Therefore analyzability and error detection of the datasets were difficult. Relational databases facilitate structured data storage and web applications open doors for platform-independent interactive visualizations. Literature has shown that supportive interactive data visualization techniques of health care records can help to uncover trends and patterns.

This thesis describes the software development and database design for the data scraping process of 26462 historical (1995 - 2016) human-readable data sets, related to allo-HSCT cell monitoring. Furthermore an interactive line chart, based on the java class C3.js, was designed, to visualize the engraftment process, as well as to identify outliers and date input errors.

As a field of application for these datasets, an explorative retrospective analysis of the naïve CD4+ T-cells, which build an important part of the adaptive immune system, were performed to test cell counts on days 30, 100, 180 and 365 after HSCT, in relation to graft material, donor's age and recipient age. Trends of inverse correlation between donor's age and count of naïve CD4+ T-cells, as well as between recipient's age and count of naïve CD4+ T-cells have been shown.

It can be expected, that this structured database and visualization represents the base for better analyzability in order to answer research questions for engraftment characteristics of patients undergoing hematopoietic stem cells transplantation at the St. Anna Children's hospital.

Keywords: data scraping, visualization, flow cytometry, HSCT, engraftment, naïve CD4+ T-cells

Kurzfassung

Das Verfahren der Durchflusszytometrie, zur Bestimmung von unterschiedlichen Erkennungsmerkmalen auf menschlichen Blutzellen ist ein etabliertes Verfahren zur Bestimmung der Zellentwicklung (Engraftment) nach hämatopoietischer Stammzelltransplantation (HSZT). Seit 1995, werden diese, auf Durchflusszytometrie basierenden, Analysen in der St. Anna Kinderkrebsforschung in Wien durchgeführt. Jedoch wurden die Messergebnisse in unstrukturierter Form, in multiplen Excel Tabellen, abgelegt, sodass die Möglichkeit einer Datenanalyse, erschwert war. Relationale Datenbanken ermöglichen eine strukturierte Ablage von Daten. Plattformunabhängige web-basierte Visualisierungstechniken erleichtern das Erkennen von Mustern und Trends in Verlaufsdaten.

In dieser Masterthesis wird die Entwicklung und Anwendung einer Software für die Datenextraktion von 26462 Datensets (von 1995 bis 2016) aus Aufzeichnungen von Durchflusszytometriedaten, sowie das Design für eine Datenbank zur strukturierten Speicherung dieser Daten, präsentiert. Ein interaktives Liniendiagramm ermöglicht den Engraftmentprozess nach HSZT darzustellen und Dateneingabefehler und Ausreißer zu erkennen.

Ein Anwendungsfall dieser Datenbank, sowie Visualisierungsmöglichkeiten wurde mit einer explorativen, retrospektiven Analyse der Anzahl von naiven CD4+ T-Zellen an den Tagen 30, 100, 180 und 365 nach HSZT und deren Zusammenhang zu Spendermaterial, Spenderalter und Empfängermaterial, präsentiert. Diese Zellen stellen einen wichtigen Teil des adaptiven Immunsystems dar. Es konnte der Trend zu einer negative Korrelation zwischen Spenderalter und Anzahl der naiven CD4+ T-Zellen, sowie zwischen Empfängeralter und naiven CD4+ T-Zellen entdeckt werden.

Zusammenfassend kann gesagt werden, dass diese strukturierte Datenbankanwendung und die darauf basierende Visualisierungstechniken eine breite Basis darstellen, um weitere Forschungsfragen über das Zell-Engraftment von PatientInnen nach HSZT im St. Anna Kinderspital zu generieren und zu beantworten.

Schlagworte: Datenextraktion, Visualisierung, Durchflusszytometrie, Stammzelltransplantation, Engraftment, naive CD4+ T-Lymphozyten

Table of Content

Preface	II
Declaration	III
Abstract	IV
Kurzfassung	V
Table of Content	VI
1 Introduction	1
1.1 Theoretical background	2
1.1.1 Hematopoiesis in humans	2
1.1.2 Hematopoietic stem cell transplantation	4
1.1.3 Fluorescence flow cytometry	5
1.1.4 Chimerism	5
1.1.5 Visualization and Interactivity	6
1.2 Problem statement	6
1.3 Work packages, research questions and state of the art analysis	8
1.3.1 Research Questions	8
1.3.2 State of the art analysis	9
2 Material and methods	11
2.1 Scope of interest	11
2.1.1 Data extraction	11
2.1.2 Statistical analysis of naïve CD4+ T-cells	13
2.2 Ethics approval	13
2.3 Development tools	13
2.4 Design of dataflow	14
2.4.1 Schematic representation of data import	14
2.4.2 Data source	15
2.4.3 Preparatory actions	16
2.4.4 Architecture of storage / database	19
2.4.5 Data scraping software	24
2.4.6 Website development and setup	28
2.4.7 Material and methods for statistical analysis	35
2.4.8 Ensuring of data quality	38

3	Results	40
3.1	Data processing	40
3.2	Interactive visualization	42
3.3	Results of statistical analysis	48
3.3.1	Web-based descriptive statistics	48
3.3.2	Normal distribution analysis	61
3.3.3	Comparison of groups	63
3.3.4	Correlations	65
4	Discussion	70
4.1	Data processing	70
4.1.1	Limitation and outlook	71
4.2	Interactive web-based visualization	71
4.2.1	Limitation and outlook	72
4.3	Statistical analyses	72
4.3.1	Limitations	75
5	Conclusion	77
	Literature	78
	List of Figures	83
	List of Tables	85
	Listings	86
	List of Abbreviations	87
	Appendix	90
A.	Ethic approval	90
B.	Declaration of consent clinical data export	93
C.	Color panel for flow cytometry	94
D.	R Script for explorative statistical analysis	95
A.	Content of attached digital medium	106

1 Introduction

Modern medicine and advanced treatments can cure rare diseases which are based on abnormalities of the blood building system (bone marrow). These disorders can be blood cancer (leukemia), malfunctions of the cellular blood components or insufficient production of blood cells, which can lead to life-threatening or fatal complications.

Since 1970, hematopoietic stem cell transplantation (HSCT), as a treatment of diseases of the blood building system, has been performed successfully on children and adolescences. HSCT is a complex procedure, where the proper blood building system, which is mainly located in the bone marrow, is replaced by the donor's blood building system. For details about HSCT see section 1.1.2. In Austria, every year around 35 pediatric patients receive an allogenic HSCT (for details see Table 1). During HSCT a regular quantification of various white blood cells of patient's blood or bone marrow takes place to monitor the hematopoiesis (see section 1.1.1) of the recipient.

Table 1: allogenic HSCT in Austria, age at HSCT <18, from 2007-2016 [1]

	bone marrow	peripheral blood stem cells	Cord Blood	Total	
				Austria	Vienna
2007	16	19	1	36	26
2008	17	13	1	31	21
2009	27	16	2	45	33
2010	18	18	0	36	26
2011	20	11	0	31	18
2012	19	19	0	38	22
2013	26	10	0	36	26
2014	27	9	1	37	25
2015	22	12	0	34	25
2016	23	16	0	39	24

Since 1996, the monitoring is performed at the Clinical Cell Biology & FACS Core Unit at the Anna Children's Cancer Research Institute (CCRI) inclusively for the St. Anna Children's hospital. The different cell populations are detected after a HSCT by characterizing their surface antigens, size and granularity using a fluorescence flow-cytometry (see section 1.1.3). First weekly and afterward monthly measurements are obligatory to monitor the process of immune reconstitution for evaluation of the clinical success of the HSCT. Generally, the reconstitution of leucocyte subpopulations takes place within the first three months after HSCT, depending on the type of cells. Usually, granulocytes are the

first detectable cells, followed by T and B lymphocytes. If all blood cells and its subpopulation are reconstituted, the engraftment process after HSCT is successful.

In 2014, Pichler et al. [2] published a retrospective study, showing that the amount of transplanted hematopoietic stem cells has no effect on engraftment of granulocytes, monocytes and erythrocytes. Further analysis of other cell types has not been performed. Detailed retrospective examinations of other cell populations are clinically important. Therefore, the engraftment characteristics of a sub-group of T-cells and the influence of graft source, recipient's and donor's age are examined in this thesis. T-cells play a central role in immune defense against foreign antigens. To achieve that goal of analysis, supportive ad hoc interactive visualization techniques of health care records can help to uncover trends and patterns [3] and are applied to find correlations and trends during the engraftment process.

This chapter gives an introduction and overview about the relevant topics. In section 1.1 the hematopoiesis, the process of regeneration of blood cells, is described, followed by an explanation of the HSCT and the methods to monitor the engraftment process. At the end of this section a short overview of the benefits of data visualization is given. Section 1.2 gives an explanation about the need of this thesis and section 1.3 provides details about the research questions and the work packages followed by a brief state-of-the art analysis.

1.1 Theoretical background

1.1.1 Hematopoiesis in humans

The most important transport medium of the human body is blood. It is a rapidly regenerating tissue and it is composed of cells and plasma [4]. Having a closer look, around 42-45% [5] of human blood consists of suspended cells in blood plasma, which also contains proteins, lipids, sugar and minerals, thus forming the non- cellular part [4, 5].

Focusing on the cellular fraction, most of the cells (99.9%) are red blood cells, also termed erythrocytes. Their main function is to transport oxygen and carbon dioxide. White blood cells form the major part of the immune system. Their dedication is to fight foreign pathogens and to learn how to fight them efficiently and effectively. Thrombocytes, also called blood platelets can form, together with components of the plasma, a hemostatic thrombus to repair the injured blood vessels to stop a potential blood loss [6]. Proportions of the different cell populations vary by sex, age [7], and by ethnic groups - or due to environmental influences.

Hematopoiesis describes the formation of the different blood cells originating from hematopoietic stem cells by cell division and differentiation, a process that mainly takes place in the human red bone marrow after birth. Classical ways of differentiation are illustrated in a simplified way in Figure 1. In 2013, Görgens et al. revised the classical and common model of the human hematopoiesis [8]. In the revised model, a multipotent progenitor cell (MPP) has self-renewing capacities and divides furthermore asymmetrically, forming a lympho-myeloid (LMPP) and an erythro-myeloid (EMP) progenitor. Cells in the LMPP pathway can eventually give rise to B-lymphocytes, T-lymphocytes, monocytes, natural killer cells and neutrophil granulocytes, whereas EMP progenitors can differentiate into thrombocytes, erythrocytes, eosinophil and basophile granulocytes (see *Figure 2*). Lymphocytes are responsible for the specific and adaptive immune response [9].

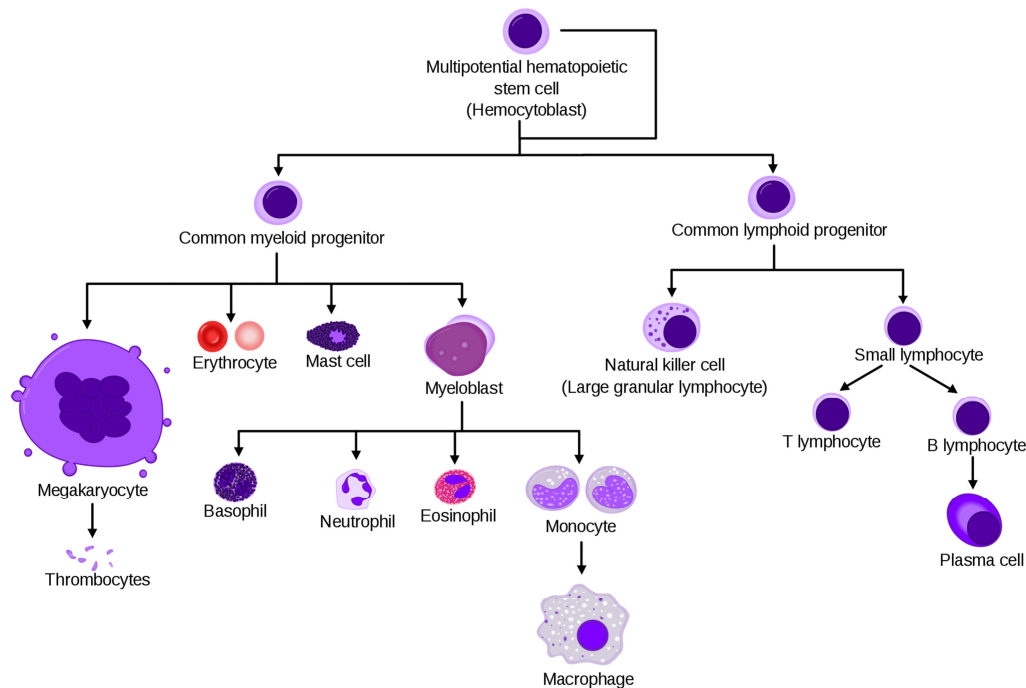


Figure 1: Classical model of the hematopoiesis in humans (adopted from [10])

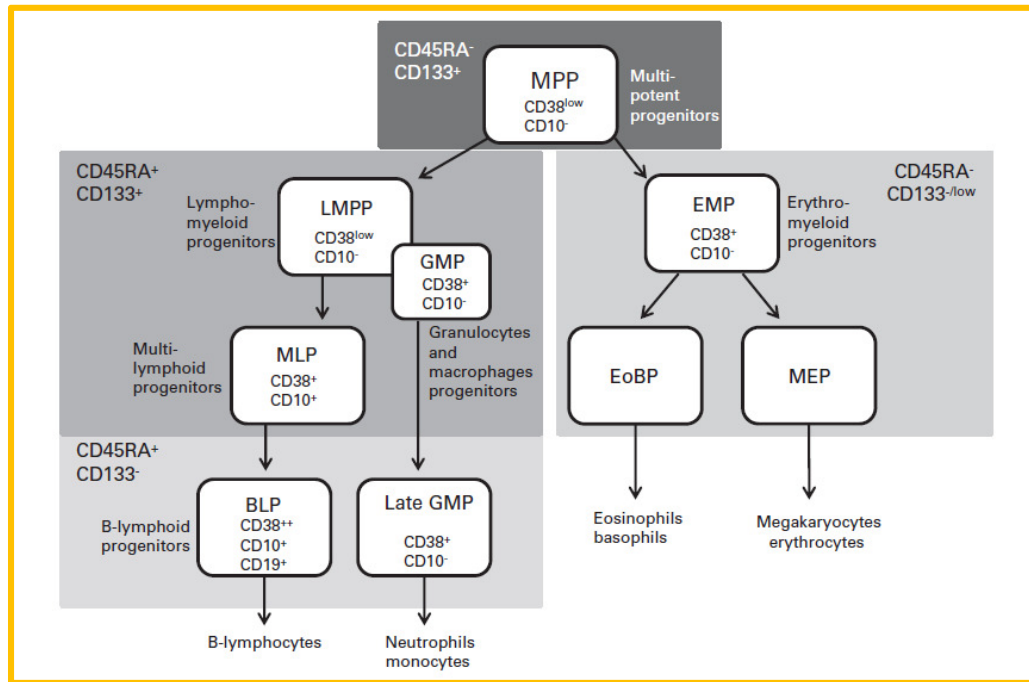


Figure 2: Revised model of the pathways of cell differentiation during hematopoiesis, adopted from [11]

The cells differ among others in terms of their function, size, and granularity, and they differ also in their specific immunophenotypic¹ surface characteristics. Some of these different cell surface molecules are standardized in the nomenclature of the cluster of differentiation (CD). For example all white blood cells have a CD45 marker, but T lymphocytes have additional CD3, the subgroup T helper cells, CD45, CD3 and CD4. CD4 positive (CD4+) T lymphocytes learn, during their maturation from naïve to effector cells, how to identify pathogens, how to support CD8+ cytotoxic T-cells, and how to kill host cells that have been infected or have undergone harmful changes [12]. Naïve T-cells are those T-cells, that have not yet encountered its cognate antigens. After antigen-recognition (priming), naïve T-cells get differentiated into effector memory T-cells with the capacity to induce a specific immune response against foreign pathogens by e.g. killing of pathogen-infected host cells [13, 14].

1.1.2 Hematopoietic stem cell transplantation

Hematopoietic² stem cell transplantation (HSCT) is a medical treatment for curing malignancies or defects of the blood-building system by the transplantation of multipotent hematopoietic stem cells (CD34+). The graft source can be bone

¹ phenotype, Greek: φαίνω *phaino* „to show“ and τύπος *typos* „type“ describes the characteristics and properties of an organism

² hematopoietic means blood building

marrow (BM), peripheral blood stem cells (PBSC) after stimulation and apheresis, or from umbilical cord blood (CB).

Depletion of CD3+ T-cells from the graft can be applied to reduce the risk of a Graft versus Host Disease (GvHD)³.

The St. Anna Children's Hospital in Vienna has performed HCST since 1980. From 01.01.1980 to 31.12.2016, 341 autologous (donor and recipient of CD34+ graft is the same person) and 717 allogenic transplantations were performed. The main patient cohort is under 18 years [15]. CD3+ T-cell depletion was performed only on PBSC grafts.

1.1.3 Fluorescence flow cytometry

The (fluorescence) flow cytometry, also known as fluorescence-activated cell sorting (FACS®⁴), is a method to detect characteristics of cells in a linear cell stream, where density and morphological properties are analyzed [16]. Dependent on specific surface antigens, cells are previously labeled by specific fluorescence antibodies. These fluorescence conjugated antibodies transmit a specific wavelength of light after activation by a laser. Therefore, cell populations can be identified.

The applied method for fluorescence flow cytometry has advanced over the years. In 1996, at the CCRI, the measurement was based on a two-color technique [17], since 2011 a ten-color measurement has been applied. Furthermore since 2002, due to the use of leucocount tubes, the sensitivity of cell detection could be approved from 100 cells / μL up to <1 cell / μL [18].

An add-on to these machines can be a sorting option, where detected cells are split-up into their populations, according to the adjustments.

1.1.4 Chimerism

The term chimerism comes from the Greek mythology and describes a creature with body parts of different animals. In medicine, the co-existence of donor's and recipient's cells in an individual, is called chimerism [19].

The aim of allogenic HSCT (allo-HSCT) is to replace the blood building system of an individual by transplantation of foreign CD34+ cells. During the engraftment process, CD34+ stem cells begin proliferation and differentiation into other cell populations (see section 1.1.1). To monitor the hematopoiesis of the foreign

³ GvHD: donors cells attack the host's tissue after allo-HSCT. The effected organs are often the skin, lungs, liver and gut.

⁴ The name FACS is registered by Becton, Dickinson and Company

immune system and the proliferation of persisting recipient's nucleated cells, two different methods can be applied. If donor and recipient are gender-unequal, fluorescence in situ hybridization (FISH) can be applied, which highlights the X and Y chromosome in two different colors. Therefore the origin of cells can be distinguished [20]. To identify the gender-independent origin of the cells, a polymerase chain reaction (PCR) method can be applied.

Chimerism monitoring is essential to be aware of a mixed chimerism, which is often associated with graft rejection or disease relapse.

1.1.5 Visualization and Interactivity

Computer-based visualization (vis) systems provide an easy-understandable language of datasets, which often consist of non-graphical content. They widen human capabilities to ask well-designed question regarding decision making [21]. Furthermore, vis opens doors to detect a tendency or a possible correlation of dataset. Another benefit of computer-based visualization, in comparison to hardcopies, is the possibility of interaction, so the end-user can adapt the appearance of representation. Moreover, a large amount of data can be displayed as a very high-level overview down through multiple ways of summarization, and datasets can be highlighted by on-mouse-over technology.

Modern web technologies allow data driven visualization in websites which can be displayed in the web browser. One of these technologies is D3.js. It is a specialized JavaScript library for bringing data into life using hypertext markup language (HTML), scalable vector graphics (SVG), and cascading style sheets (CSS). C3.js⁵ [22] is a further development basing on D3.js. It allows the user to customize graphs without writing the D3.js code. The webpage <http://c3js.org/examples.html> provides many examples and illustrates cases of C3.js.

1.2 Problem statement

Starting in July 1995, at the laboratory of Clinical Cell Biology and FACS® Core Unit at the CCRI specimens were analyzed for patients at the St. Anna Children's hospital, for cell engraftment documentation. The documentation was essential for the pediatricians and medical experts at the HSCT ward, to adjust the care during the treatment of HSCT. In 1995 only seven subgroups of white blood cells (leucocytes) could be analyzed. The delectable cells, beside the leucocytes, were T-cells (CD3+), T helper cells (CD4), cytotoxic T-cells (CD8), natural killer cells

⁵ www.c3js.org D3-based reusable chart library

(NK), monocytes (Mono), granulocytes (Granulo) and hematopoietic stem cells (CD34).

Over time, using new markers and labels, more cell types were detected by flow cytometry and reported to the clinic. These additional cells were natural killer T cells (NKT), B-cells (CD19), activated CD3+ cells (38+ of CD3), naïve CD4+ T-cells, naïve CD8+ T-cells, B precursor cells, eosinophil granulocytes (Eos), $\gamma\delta$ T cells (g-d), normocytes⁶, basophil granulocytes and granulocytic myeloid-derived suppressor cells (CD33^{dim}).

The written reports were Microsoft Excel tables with numeric results and remarks (for an example see Figure 3 and Figure 4), which were sent by fax to the requesting clinical ward. For every patient, an Excel table was created after allo-HSCT. This way of documentation started in July 1995 and the structure of the tables varied over years. Figure 4 illustrates, that in fields, where numbers are expected, text was inserted and instead of new columns additional test results became part of the remarks (german: "Bemerkungen"). Therefore, working with these issues became a problem for analysis and comparison of different patients' documentation.

Additionally, because of a long list of values, errors of data insertion and outliers were retrospectively hard to detect. Error proofing was not part of these records in Excel. A fast visualization of cell engraftment progression was not available to find patterns and trends of cell engraftment. Therefore, for investigations and studies, based on that data, manual seeking in multiple files was essential.

⁶ in this content, normocyte is a nucleated preliminary stage of an erythrocyte and can, generally, exclusively be found in the bone marrow

	A	B	C	D	E	F	G	H	I	J
1	Nachname V									
2	alle Angaben als Zellen pro µl Blut									
3										
4	Datum	Tag nach	Leuko	CD3	CD4	CD8	NK	Mono	Granulo	CD34
5		Transpl.								
6										
7	08.03.1996	-4	400	48	12	20	130	36	120	0,2
8	11.03.1996	-1	200	1	0,4	0,6	100			0,3
9	12.03.1996	0	250	9	6	3	78	23	130	0
10	13.03.1996	1	500	55	30	18	250	30	120	0,8
11	14.03.1996	2	400	48	27	16	176	17	96	0,7
12	15.03.1996	3	500	80	30	20	245	29	95	0,7
13	18.03.1996	6	500	135	40	80	210	32	90	1,3
14	19.03.1996	7	850	264	72	153	374	21	119	1,3
15	21.03.1996	9	850	213	69	136	366	88	123	0,1
16	22.03.1996	10	950	200	57	76	314	71	171	0,7
17	26.03.1996	14	1050	184	121	63	231	137	420	0,8
18	28.03.1996	16	1850	250	85	96	342	425	740	1,3
19	29.03.1996	17	2000	360	180	130	270	420	1000	2
20	01.04.1996	20	2100	315	149	109	143	210	1407	1,7
21	04.04.1996	23	3100	434	177	96	78	329	2170	1,1
22	09.04.1996	28	4450	258	129	53	67	378	3694	0,9
23										
24										

Figure 3: first (year 1996) flow cytometry report of the CCRI, the header just consists of the name of the patient.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U						
1	Nachname Vorname (01.1.95)				HR ALL		34 kg, ,		transpl. am X.X.XX mit KM, (DE, NNNNNNN,w)																		
2							2. TX		37 kg, ,		transpl. am X.XX.XX mit KM, (DE-XX00000,w)																
3	alle Angaben als Zellen pro µl Blut (Ausnahme Angaben in %)												38+ in %CD3		FA+62L (naive 4)		FA+62L (naive 8)		Bemerkungen								
4	Datum	Tag nach	Leuko	CD3	CD4	CD8	NK	Zyto T	Mono	Granulo	CD34	CD19	CD38+ (%CD3)	RA/27/31 (%CD4)	RA/27/31 (%CD8)	B Vorl (%34)	Eo/µl (%CD3)	g-d (%NC)	Normo Baso	neu seit 7/14 Bemerkungen							
5																											
6																											
7																											
8	4.8.08	-1	*93	0			0		0	90	0																
9	5.8.08	0 vor	*8	0			0		0	8	0																
10	5.8.08				Patient erhält 3.2 myeloide CD34 & 27 CD3 (x10E6 kg)																						
11	6.8.08	1	*109	0,1	Mono & NK = 1					105	0,03																
12	12.8.08	7	*3,5	0			?		0,3	2	0	?									"lymphoide" Zellen ca 1/µl						
13	18.8.08	13	*1	nein			ja ?		ja ?	nein ?																	
14	19.8.08	14	*2	ja ?			ja ?		ja ?	?																	
15	22.8.08	17	*23	1			5	0,1	5	10	0,02	0	91								N						
16	25.8.08	20	*222	6	5	1	18	0	50	140	0,04	0,02 ?	93								NC unsortiert = 100%w						
17	26.8.08	21	500	20			60	0	40	350	0,3	0	89														
18	19.08	27	1.300	17	12	4	45	0	370	815	2,5	0	78														
19	1.9.08				100%w	100%w	100%w		100%w	100%w	94/96=										sort FISH						
20	2.9.08	28	1.200	15	10	3	60	0,3	360	710	2,5	0	72	0	0						6% d CD3 = g-d						
21	2.9.08		7.100	20	8	8	90	2	305	4700	60	10	87								KM, + 16% Normo; 11% d CD3=g-d; (Eo=280/µl)						
22	2.9.08			0/1=w			100%w		100%w	100%w	100%w		sort Normo = 100%w								sort FISH 87% d 34 = B Vorl						
23	8.9.08	34	2.300	65	28	35	170	2	430	1540	0	2	88	0	0						2% d CD3 = g-d						
24	8.9.08				100%w	100%w	100%w		100%w	100%w	68/71=										sort FISH						
25	10.9.08	36	3.200	65	29	34	185	2	625	2175	2	0									2% d CD3=g-d; (Eo=550/µl)						
26	15.9.08	41	5.100	100	36	59	415	2	840	3545	3,5	1	83	0	2 ?						2% d CD3=g-d; (Eo=1170/µl); 1/4 d 34=B Vorl						
27	23.9.08	49	4.500	55	30	20	235	0	650	3350	1,8	18	74	0	0 ?						4% d CD3=g-d; (Eo=585/µl); 20% d CD34 = B Vorl						
28	23.9.08				100%w	100%w	100%w		100%w	100%w	100%w	100%w									sort FISH						
29	3.10.08	59	3700	60	35	21	290	1	535	2645	0,7	65	64	0	4 ?						5% d CD3=g-d; (Eo=450/µl)						
30	3.10.08				100%w	100%w	100%w		100%w	100%w	100%w	100%w									sort FISH						
31	3.10.08		5900	35	20	13	165	2	295	4600	205	470	64	0	8 ?						KM, + 5% Normo; 3% d CD3=g-d; (Eo=290/µl)						
32	3.10.08				100%w	100%w	100%w		100%w	100%w	100%w	100%w	sort Normo = 100%w								sort FISH 88% d 34 = B Vorl						

Figure 4: flow cytometry report of the CCRI, containing a header with the corresponding data of the patient and information about the transplantation. New information (see row 20) was just added in the corresponding field in column P ("Bemerkungen").

1.3 Work packages, research questions and state of the art analysis

1.3.1 Research Questions

In this thesis, the following work packages (WP) and research questions (RQ) are addressed:

- WP1 Feasibility: Transfer of existing records of flow cytometry analyses data into a platform, which allows an easier data access and advanced analyzability.
- WP2 Feasibility: Development of a web-based interactive visualization tool for flow cytometry data with the option to filter results according to graft source, human leukocyte antigen (HLA) mismatch and number of transplanted CD34+ and CD3+ cells.
- RQ1: Do graft source, age of donor and age of recipient influence the amount of naïve CD4 T cells on days 30,100,180,365 after allo-HSCT?
 - RQ1.1 Statistical Analysis: Does the graft source (BM, PBSC, CD3 depleted PBSC) influence the engraftment process (absolute cell count) of naïve CD4+ T-cells after allo-HSCT?
 - RQ1.2 Statistical Analysis: Does the engraftment process (absolute count) of naïve CD4+ T-cells correlate with the age of the donor?
 - RQ1.3. Statistical Analysis: Does the engraftment process (absolute count) of naïve CD4+ T-cells correlate with the age of the recipient?
- WP3 Feasibility: Development of a web-based tool for descriptive statistics and distribution of data for questions RQ1.1-RQ1.3.

1.3.2 State of the art analysis

Rind et al. surveyed in their scientific paper “Interactive Information Visualization to Explore and Query Electronic Health Records” [23] different electronic health record systems. Fourteen state-of-the-art visualization systems were reviewed in detail. They described the demand of visualization tools for clinical researchers to detect unknown patterns, thus giving the chance to discover surprising outcomes and the support for clinical decision making.

Regarding immune reconstruction after allo-HSCT, delayed immune reconstruction is a reason for morbidity and mortality after HSCT. Naïve CD4+ T-cells play an important role host defense via identification and defense of newly encountered antigens [24, 25]. Goldberg et al. [25] illustrated the pattern and timings of the engraftment process of T-cell subsets. In their retrospective data analysis of 375 recipients post HSCT after T-cell depleted allogeneic graft (median age 40 (2-68)), they concluded that, naïve CD4+ T-cell engraftment is not robust in the first 6 month and deficiencies in CD3+ and in particular in CD4+ T-cell

engraftment correlate with an increased risk of infections [26, 27]. Regarding naïve CD4⁺ T-cell engraftment, they showed that recipient's age did not strongly effect the reconstitution in their data set, but in murine HSCT studies the donor's age does [24]. Furthermore, Servais et al. [28] showed no significant difference in the number of naïve CD4⁺ T-cells between 30 recipients of umbilical cord blood and 36 recipients of mismatched unrelated donor graft (source: 27 PBSC, 9 BM).

Additionally, Azuma et al. [24] and Hirayama et al. demonstrated in their murine HSCT studies, that the number of naïve CD4⁺ T-cells correlated inversely with donor age and that generally naïve CD4⁺ T-cells from younger donors showed a significantly higher proliferation rate compared to adult donors. Naïve CD4⁺ T-cells had to follow the thymus-dependent pathway and increase rapidly in younger recipients and when transplanted from young donors.

The explorative, retrospective analysis of the influence of graft source, recipient's age and donor's age for the engraftment process of naïve CD4⁺ T-cells after allo-HSCT at patients at the St. Anna Children's Hospital of Vienna is part of this thesis.

2 Material and methods

This chapter deals with the material (software and development tools) and methods used in this thesis. It describes the scope of investigation, the relevant data, the including and excluding criteria for the datasets used in the investigation, and the technical approach leading to the process of visualization and analysis of the datasets.

The legal framework, including approved ethics can also be found in this chapter.

2.1 Scope of interest

This section is divided into two parts. The first part describes the methodology of how the existing flow-cytometry-based analysis data were checked, extracted and finally saved into a relational data base. The second part covers the scope of investigation and the methodology for the explorative statistical analysis of naïve CD4 T-cells.

2.1.1 Data extraction

At the department of documentation at the St. Anna Children's hospital, clinical data about all patients and their autologous and allogenic HSCT is stored in an Excel table, named `STAMM1C.xls`. It contains, inter alia, information about the patient, the preparation and conditioning plan and information about the donor and events during transplantation. The file `DLINACHK.xls` contains information about dates of additional donor lymphocyte cell infusions (DLI). Selective information of the clinical documentation of these two Excel files, according to the ethical approval (see section 2.2), was provided by Eva Nowak, the assistant of the transplantation ward at the St. Anna Children's hospital. A declaration of consent, signed by the head of the stem cell transplantation ward at the St. Anna children's hospital, with the permission to use a part of the clinical documentation, can be found in Appendix B. The following data of patients were included in the data extraction process:

Source

Documentation table STAMM1C.xls

Data:

- TX (number of transplantation)
- TR_NR (transplantation per patient)
- FAMILIENNAME (last name of patient)
- VOR_NAME (first name of patient)
- GEB_DAT (date of birth of patient)
- DIAG_CODE + remark (code of disease)
- ERST_DIAG_TXT (detailed description of disease)
- KMT_DAT (date of HSCT)
- KG_KMT (body weight of patient at HSCT)
- KMT_ART (type of graft's origin [autologous/allogenic])
- SZ-ART (graft's source [BM/PBSC/Cord blood (CB)])
- S_GE (gender of donor)
- S-GRAD (relationship donor/recipient)
- S_GEBDAT (date of birth of donor)
- CD34kg ($CD34+ \cdot 10^6$ cells/kgBW⁷ in graft material)
- CD3kg ($CD3+ \cdot 10^6$ cells/kgBW in graft material)
- ZAHL MM (HLA mismatch donor/recipient)

Source:

Documentation table DLINACHK.xls

Data:

- Names of patients who received additional DLI after HSCT
- Date of DLI

Furthermore, the historical documentation for fluorescence flow cytometry data, which is stored at the CCRI network path G:\FORSCHUNG\LAB1\excel\allogen, was analyzed. Historically, for every patient a new file (in total 431 separate files) had been created. To avoid an interruption of the daily routine a local copy of all files was done on 13/12/2016.

Therefore, records, starting at the first performed allo-HSCT (25/10/1980) to the transplantation on the 13/12/2016, are the scope of this investigation. Both data sources (documentation table STAMM1C.xls and documentation table DLINACHK.xls) are the basis for the structure of the database.

⁷ kilogram body weight

2.1.2 Statistical analysis of naïve CD4+ T-cells

Based on the stored data a database query could be performed. Thus, the values of naïve CD4+ T-cells could be selected and built part of the dataset for statistical analysis.

Records, in the timespan from 22.11.2000 to 05.12.2016 built the scope of interest for statistical analysis and the basis for answering the research questions in section 1.3.1. Since 22.11.2000 a fluorescence marker for surface antigen CD62L was installed for fluorescence flow cytometry to detect naïve lymphocytes. Because of lack of change reports at this period of time, the date of analysis change was obtained by comparison of the hardcopy daily records in the historical archive.

2.2 Ethics approval

Ethics approval, for working with sensitive, personalized clinical data in a retrospective data analysis, was applied at the Ethics Committee of the Medical University of Vienna⁸ on 01/12/2016, an update on required on 02/12/2016. A final positive vote was given by the committee on the 26/01/2017 (see Figure 5).

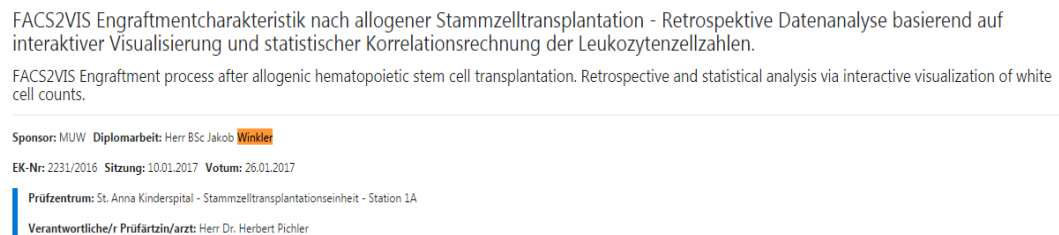


Figure 5: Positive vote for the ethics approval for the data analysis of the project for the master thesis [29].

The entire ethic approval document can be found in Appendix A.

2.3 Development tools

In Table 2 the software products used to establish this thesis and its related work, including program development and statistical calculation, are listed.

⁸ <http://ethikkommission.meduniwien.ac.at/>

Table 2: Required software for the master thesis and the related work

Name of Application	Version	Scope
JetBrains Webstorm	2016.3.4	php/js
Microsoft Visual Studio Community 2015	14.0.25431.01 Update 3	C# development of "pqparse_xls" and "patient2db"
ORACLE MySQL Workbench	6.3.8	Database development Database content management
https://regex101.com/ (online)	(accessed 01/2017)	Testing and development of regular expressions
R	3.4.0 (2017-04-21)	statistical testing
XAMPP	3.2.2.	Apache, MySQL
MS Word 2010 Professional Plus	14.0.7.182.5000	Writing of master thesis

2.4 Design of dataflow

2.4.1 Schematic representation of data import

The data flow of the information transport from the historical information of the Excel sheets into the relational database is shown in Figure 6. The orange-framed boxes describe the origin of the data; the green boxes the preparatory tasks, which had to be fulfilled manually; and the black-framed boxes the C# applications to parse out the relevant information of the Excel files. The light blue framed boxes describe the final destination where the extracted data is stored. The yellow box is a representation of a web application to add data directly to the database. A detailed description of the different elements can be found in the following sections.

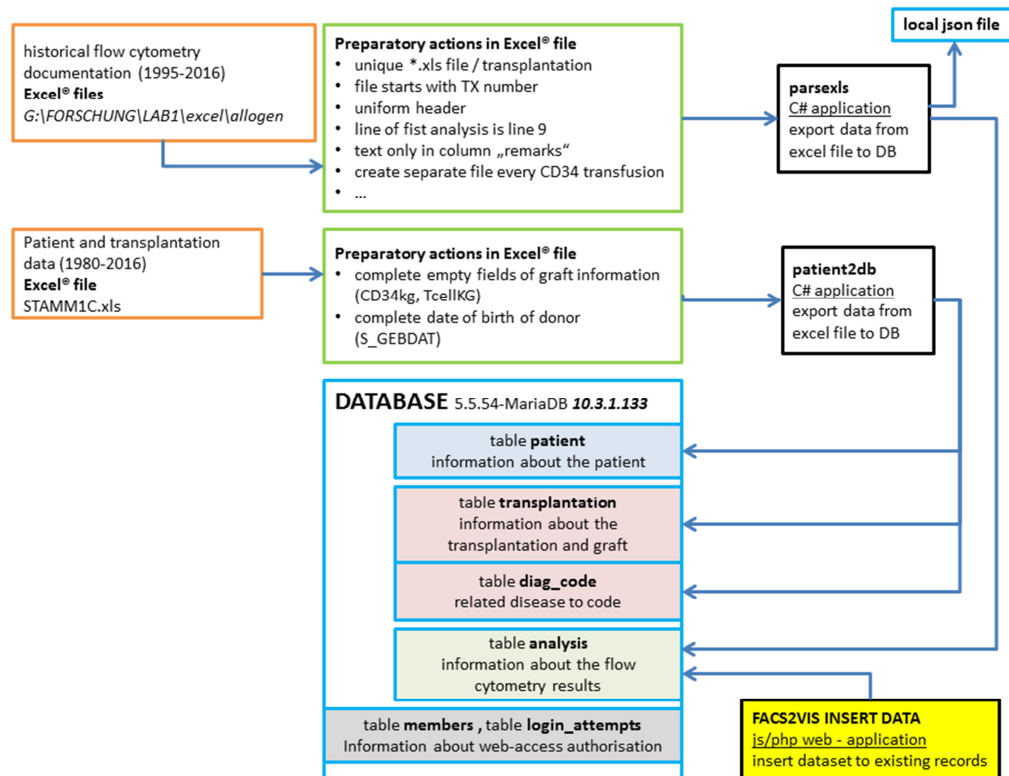


Figure 6: Overview of data flow (content of excel files into database)

2.4.2 Data source

The documentation of flow cytometry has been stored at the CCRI network at G:\FORSCHUNG\LAB1\excel\allogen\, since 1995. In this directory different sub-directories could be found. Gerhard Fritsch, PhD, the head of the department of Clinical Cell Biology & FACS Core Unit at the CCRI, stored the obtained flow cytometry-based results in this directory, including at subdirectories like “old patients”, and “deceased patients”. All files were copied to an entire directory and moved to the G:\FORSCHUNG\FACS2VIS\unbearbeitet (engl. unprocessed) to avoid interruption of the daily routine. This local copy was done on 13/12/2016. In this folder the directory “bearbeitet” (eng. processed) was created to separate files which had passed preparatory actions and processing. Files after preparatory actions were created in the directory G:\FORSCHUNG\FACS2VIS\bearbeitet, within the directory “inDBeingelese” (eng. “Stored in Database”), which contains the Excel files that were already parsed as well as content stored in the database.

The pre-selected Excel file “STAMM1C.xls”, which contains relevant information about patient and transplantation, was sent to jakob.winkler@ccri.at by Ewa Nowak (ewa.nowak@ccri.at).

2.4.3 Preparatory actions

An automatized parsing routine within the C# applications needs preparatory actions to implement methods for data extraction. For example, these preliminary actions should be standardization of column order, a standardized description of method of analyses. Additionally, remarks and comments must be placed explosively in the column for “remarks”.

2.4.3.1 Preparatory actions in Excel files, containing flow cytometry results

This task was partially outsourced. Chronologically, the newest 77 records were analyzed by Jakob Winkler. The remaining 354 records were analyzed by the following members of the department: Dieter Printz, Dijana Trbojevic, Julia Stemberger and Angela Halfmann. This preparatory task was performed following a guideline (see Figure 7) by Jakob Winkler and under his supervision.

[illegible]

Figure 7: Guideline how to format patient's flow cytometry report to be extracted correctly

Filename

The filename consists of the transplantation number of the STAMM1C.xls following the proper number of transplantation, then the last name and first name, ending up with the extension xls (e.g.: 888_1_Winkler_Jakob.xls).

Of note, each transplantation or CD34+ boost (infusion of CD34+ stem cells) counts as a separate transplantation independent of whether a single patient was transplanted once or more often.

Standardized order of column names

The names of the header and the corresponding content must be in the following order: "Datum" (engl. date), "Tag nach" (engl. day after), "Leuko" (total count of white blood cells), "CD3" (T-cells), "CD4" (T-helper cells), "CD8" (cytotoxic T-cells), "NK" (natural killer cells), "ZytoT" (natural killer T cells), "Mono" (monocytes), "Granulo" (granulocytes), "CD34" (hematopoietic stem cells), "CD19" (B-lymphocytes), "CD38+" (activated T cells), "naive 4" (naïve CD4), "naive 8" (naïve CD8), "B Vorl" (progenitors of B-lymphocytes), Eo\µl (eosinophil granulocytes), "g-d" (γδ-T cells), "Normos" (normocytes), "Bemerkungen" (engl. remarks). This order is essential, because the parsing algorithm is based on the column number; therefore, the first 20 columns must contain the content in the described order.

First row of analysis data

The first row of analysis data must be row 9. The parsing algorithm starts parsing at row 9 and ignores content above.

Remarks

Remarks and additional text information must be placed in the column "Bemerkungen". Due to A4 page size restriction, text information was placed inconsistently in different cells. If a remark was placed on a separate row with the same date of analysis, the row should be deleted and the content placed in the column "remark" with the same date of analysis.

Detailed information about test method

In general, all values within single fields of the Excel table represent flow cytometry analyses of peripheral blood cells, based on the unit cells/µl. In contrast, "g-d" and "CD38+" are shown in % of "CD3", "naive 4" in % of "CD4", "naive 8" in % of "CD4" and "B Vorl" in % of "CD34". The value of "Normo" is also in % of all nucleated cells (NC). NCs consist of leucocytes and normocytes. Some datasets are analyses of different material (bone marrow (BM), pleural punctate, liquor). Also the analysis method can be different. Chimerism analysis results are placed in %. If the underlying analysis method is a PCR, the result is

NN%⁹d (percentage of donor cells) or NN%⁹r (percentage of recipient cells); in contrast if the underlying method is FISH the result can be NN%w (percentage of female cells) or NN%m (percentage of male cells). Proportions with absolute numbers (e.g. 46/49w) had to be calculated and replaced with the new value (e.g. 94%w).

To analyze the chimerism of specific cell types such as NK cells or T-cells etc., appropriate cells have to be previously sorted. Dependent on the method used, the abbreviations “sort-PCR” or “sort-FISH” have been written into the field “remark”. These comments about analysis material and method have to be checked and possibly supplemented.

Additional information and columns

Measurement results, which were listed in the column of “remarks” (e.g.: 0.5% d CD3=g-d; (Eo= 70/μl)), were mapped to the according column.

Further analysis and additional columns could be cleaned after consultation with the head of the laboratory, Gerhard Fritsch, PhD.

2.4.3.2 Preparatory actions in Excel files, containing information about patient and transplantation

The file STAMM1C.xls, containing relevant information about patient and transplantation, had some data gaps. Empty Excel-fields could be found in the column of “S_GEBDAT” (birthday of donor), “CD34kg” (CD34+ ·10⁶ cells/kgBW¹⁰ in graft material) and “TcellKG” (CD3+ ·10⁶ cells/kgBW in graft material). The gap of missing graft-cell information could be closed by comparison with the original hardcopy reports in the archive. If available, only the amount of myeloid stem cells was recorded after consultation with Gerhard Fritsch, PhD and Susanne Matthes-Leodolter, MD. Donor birthday information, if missing at the St. Anna Children’s hospital archive, was requested by Ewa Nowak from the Austrian bone marrow donor registry. If only donor’s age at the time of transplantation was known, the first of January of the corresponding year was quoted.

2.4.4 Architecture of storage / database

The following section describes the setup and design of database. In December 2016 the database structure was developed on the personal computer with the support of XAMPP and ORACLE MySQL Workbench. In February 2016 the database was moved to the “synology-ha” server at the CCRI network to ensure daily backups and facilitate database access, including restrictions, from all PCs

⁹ NN is a variable for numbers from 0 to 100

¹⁰ kilogram body weight

in the network group. The access to the database was provided by Lukas Schneider, head of the IT department.

type of database: 5.5.5.54-MariaDB (open source database)

server IP address: 10.3.1.133

database name: facs_db

username: facs2vis

2.4.4.1 Access control

The database is protected by username and password and the logging is activated. Furthermore, web-access to pages which communicate with the database is protected by a login and access control routine. This procedure was adapted from wikiHow [30].

2.4.4.2 Software considerations

For facs_db database table setup, the graphical user interface (GUI) of MySQL workbench was used. This software provides full database access supported by a GUI, additionally to a command line interface to execute Structured Query Language (SQL) statements. Furthermore entity relationship (ER) models and ER diagrams can be created to visualize relationships and dependencies between the tables and database entries.

Table “analysis”

The table `analysis` contains the columns including values analyzed cell counts. “id_analysis” is the primary key for this table. “id_transplantation” and “id_analysis” are the foreign keys in relation to tables `patient` and `transplantation`, because analysis datasets are exclusively valid if assigned to a patient and a transplantation. “anl_date” contains definition of specimen: peripheral blood (PB), bone marrow (KM), bronchial alveolar lavage (BAL), Liquor, pericardial punctate, pleura punctate (germ. “Pleurapunktat”), tracheal secretion (germ. “Trachealsekret”). Details see in Table 3.

Table 3: Table “analysis” of database “facs_db”. “Field” describes the name of the column, “Type” the datatype of the variable, “Null”, if content is obligatory, “Key” if field is assigned to be a unique primary key or a foreign key (MUL). “Content/Remark” describes the content of the field.

Field	Type	Null	Key	Content/Remark
id_analysis	int(10) unsigned	NO	PRI	auto_increment
id_transplantation	int(10) unsigned	NO	MUL	refers to table “transplantation”
id_patient	int(10) unsigned	NO	MUL	refers to table “patient”
anl_date	date	NO		date format: YYYY-MM-DD
anl_material	varchar(15)	YES		PB, KM, BAL, liquor...
anl_method	varchar(45)	YES		FACS, sort-pcr, sort-facs
anl_WBC	float	YES		count of white blood cells / % chimerism
anl_WBC_sort	varchar(10)	YES		chimerism differentiator (d/r), (m/w)
anl_WBC_leucocount	tinyint(1)	YES		true(1) if anl_WBC counted with “leuco count”
anl_CD3	float	YES		count of CD3+ T-cells / % chimerism
anl_CD3_sort	varchar(10)	YES		chimerism differentiator (d/r), (m/w)
anl_CD4	float	YES		count of CD4+ T-cells / % chimerism
anl_CD4_sort	varchar(10)	YES		chimerism differentiator (d/r), (m/w)
anl_CD8	float	YES		count of CD8+ T-cells / % chimerism
anl_CD8_sort	varchar(10)	YES		chimerism differentiator (d/r), (m/w)
anl_NK	float	YES		count of NK cells / % chimerism
anl_NK_sort	varchar(10)	YES		chimerism differentiator (d/r), (m/w)
anl_NK56neg	float	YES		count of NK56neg cells / % chimerism
anl_NK56neg_sort	varchar(10)	YES		chimerism differentiator (d/r), (m/w)
anl_NKT	float	YES		count of NKT cells / % chimerism
anl_NKT_sort	varchar(10)	YES		chimerism differentiator (d/r), (m/w)
anl_Mono	float	YES		count of monocytes / % chimerism
anl_Mono_sort	varchar(10)	YES		chimerism differentiator (d/r), (m/w)
anl_Granulo	float	YES		count of granulocytes / % chimerism
anl_Granulo_sort	varchar(10)	YES		chimerism differentiator (d/r), (m/w)
anl_CD34	float	YES		count of CD34+ cells / % chimerism
anl_CD34_sort	varchar(10)	YES		chimerism differentiator (d/r), (m/w)
anl_CD19	float	YES		count of CD19+ cells / % chimerism
anl_CD19_sort	varchar(10)	YES		chimerism differentiator (d/r), (m/w)
anl_CD38	float	YES		percentage of CD38+ cells (%CD3+) / % chimerism
anl_CD38_sort	varchar(10)	YES		chimerism differentiator (d/r), (m/w)
anl_CD4_naiv	float	YES		percentage of naïve CD4 T-cells (%CD4+) / % chimerism
anl_CD4_naiv_sort	varchar(10)	YES		chimerism differentiator (d/r), (m/w)
anl_CD8_naiv	float	YES		percentage of naïve CD8 T-cells (%CD8+) / % chimerism
anl_CD8_naiv_sort	varchar(10)	YES		chimerism differentiator (d/r), (m/w)
anl_B_precursor	float	YES		percentage of B precursor cells (%CD34+) / % chimerism
anl_B_precursor_sort	varchar(10)	YES		chimerism differentiator (d/r), (m/w)
anl_Eos	float	YES		count of eosinophil granulocytes / % chimerism
anl_Eos_sort	varchar(10)	YES		chimerism differentiator (d/r), (m/w)
anl_gammadelta	float	YES		percentage $\gamma\delta$ T cells (%CD3+) / % chimerism
anl_gammadelta_sort	varchar(10)	YES		chimerism differentiator (d/r), (m/w)
anl_normo	float	YES		percentage normocytes (%NC) / % chimerism
anl_normo_sort	varchar(10)	YES		chimerism differentiator (d/r), (m/w)
anl_Baso	float	YES		count of basophil granulocytes / % chimerism
anl_Baso_sort	varchar(10)	YES		chimerism differentiator (d/r), (m/w)
anl_CD33dim	float	YES		count of CD33 ^{dim} / % chimerism
anl_CD33dim_sort	varchar(10)	YES		chimerism differentiator (d/r), (m/w)
anl_CD3neg7pos16neg56pos	float	YES		count of CD3 ⁷⁺ 16 ⁵⁶⁺ cells / % chimerism
anl_CD3neg7pos16neg56pos_sort	varchar(10)	YES		chimerism differentiator (d/r), (m/w)
anl_remark	varchar(250)	YES		additional information / remark

Table “patient”

Table “patient” contains a unique continuous number of the patient. This number has no relation to numbers of the clinical records of the St. Anna Children’s hospital. The datasets with first name, family name and patient’s birthday were adopted from the clinical documentation “STAMM1C.xls”. Details of structure see Table 4.

Table 4: Table “patient” of database “facs_db”

Field	Type	Null	Key	Content/Remark
id_patient	int(10) unsigned	NO	PRI	auto_increment
pat_fristname	varchar(20)	NO		patient's first name
pat_lastnamename	varchar(20)	NO		patient's last name
pat_birthday	Date	NO		date format: YYYY-MM-DD

Table “transplantation”

Table “transplantation” contains the number of transplantation adopted from the clinical documentation STAMM1C.xls. This number is unique and serves as the primary key. Further data fields are designed according to their content in STAMM1C.xls, and the foreign key “pat_diseaseCode” is related to the table “diag_code”, where every code has its proper description. The field “pat_diseaseText” contains detailed information about the disease additional to the categorical and classified field “pat_diseaseCode”. For details of structure see Table 5.

Table 5: Table “transplantation” of database “facs_db”

Field	Type	Null	Key	Content/Remark
id_transplantation	int(10) unsigned	NO	PRI	equal to column “TX” in STAMM1C.xls
id_patient	int(10) unsigned	NO	MUL	refers to table “patient”
pat_diseaseCode	int(11)	NO	MUL	refers to table “diag_code”
pat_diseaseText	varchar(20)	YES		detailed description of disease
kg_kmt	float	YES		patient's bodyweight on the day of HSCT
dateOfTx	date	NO		date of HSCT, date format: YYYY-MM-DD
tx_material	int(10)	NO	MUL	Classification according “Stamm1C” 1=BM, 2=PBSC
tx_number	varchar(20)	YES		number of graft, stem cell register (currently not in use)
tx_sex	tinyint(1)	NO		gender of donor (0 = female, 1 = male)
tx_relationship	varchar(20)	NO		degree of relationship (donor to recipient)
tx_CD34_kg	float	YES		CD34+ ·10 ⁶ cells/kgBW in graft material
tx_CD3_kg	float	YES		CD3+ ·10 ⁶ cells/kgBW in graft material
tx_donor_MM	int(11)	YES		count of HLA mismatches of 10
tx_donor_birthday	date	NO		donor’s date of birth, date format: YYYY-MM-DD
tx_CD3_AB	tinyint(1)	NO		CD3 graft depletion (currently not in use)
tx_PtCy	tinyint(1)	NO		patient received post-HSCT cyclophosphamide? (currently not in use)
DLI	tinyint(1)	NO		default value: 0, 1 if patient received DLI

Table “diag_code”

According to STAMM1C.xls this table contains the mapping rules for the disease diagnosis code. Details of structure see Table 6.

Table 6: Table “diag_code” of database “facs_db”

Field	Type	Null	Key	Content/Remark
id_diag_code	int(11)	NO	PRI	equal to column “DIAG_CODE” in STAMM1C.xls
diag_name	Varchar(200)	NO		name of disease

The following diseases are registered in the table: acute lymphatic leukemia (ALL), acute myeloid leukemia (AML), chronic myeloid leukemia (CML), neuroblastoma (NBL), severe aplastic anemia (SAA), congenital anemia, Non-Hodgkin-Lymphoma, Ewing tumor, storage diseases, rhabdomyosarcoma (RMS), brain tumor, immune defect, malign Langerhans cell histiocytosis (LCH), Huntington's disease (HD), familial hemophagocytic lymphohistiocytosis, myelodysplastic syndrome (MDS), autoimmune diseases, chronic myelomonocytic leukemia (CMML), dyserythropoetic anemia, thalassemia, sickle cell anemia, osteosarcoma and further tumors.

Table “HSC_Source”

According to STAMM1C.xls this table contains the mapping rules for the material description of the hematopoietic stem cell (HSC) source, which includes either bone marrow, cord blood (CB) or stem cell enriched peripheral blood. For the latter, patients were treated first with the growth factor GM-CSF to flush out stem cells from bone marrow into the peripheral blood, a process known as mobilization. Thereafter, cells were harvested via a blood leukapheresis. Some patients got a combination of different stem cell sources. The classification is according to STAMM1C.xls, table ‘SZ-ART’ (1=bone marrow, 2=peripheral blood stem cells, 3=BM and PBSC, 4=cord blood, 5=BM and CB, 6=SC and CB). Details of structure see Table 7.

Table 7: Table “HSC_Source” of database “facs_db”

Field	Type	Null	Key	Content/Remark
id_HSC_Source	int(11)	NO	PRI	equal to column “SZ-ART” in STAMM1C.xls
HSC_Source_nameshort	varchar(10)	NO		abbreviation of graft source
HSC_Source_namelong	varchar(45)	NO		name of graft source

Table “login_attempts” and table “members”

These two tables are part of the protected login and registration process to access protected websites of this project. The structure is adopted following the guidelines in [30].

2.4.5 Data scraping software

This section describes the extraction process of flow cytometry data located in the historical records (file format: Excel) into the database.

The graphical development for this part was Microsoft Visual Studio Community 2015 (details see section 2.3). Both applications are developed as a console application. To access the information stored in Excel tables, the Microsoft interface OLE DB¹¹ [31] was used. A graphical user interface was not implemented, because the applications serve exclusively for the developer, Jakob Winkler, to extract data from Excel tables. Thus, there was no need for other users to work with the application.

The entire code of these applications can be found on the attached disc in the directory “development”.

2.4.5.1 *patient2db*

“Patient2db” is a console application, written in C# to connect to the documentation file “Stamm1Ccorr_final.xlsx”. “Stamm1Ccorr_final.xlsx” contains selective columns of the original file “STAMM1C.xls” regarding information about patient and transplantation. This data can be copied to the application presented in this master project. A written consent is attached in Appendix B.

“Patient2db” is developed as an object oriented approach which contains the class `facMySQL` including the methods `facMySQL()`, `datasetPatient2db()` and `datasetTx2db()`. The main program is set up as follows (flow chart see Figure 8): In the first section the object `mySQL` is built according to the default constructor of the class `facMySQL`. The following section sets up the OLEDB connection to the input file source located at `H:\Master\Uebersetzungstabelle\STAMM1Ccorr_final.xlsx`.

Then, an algorithm extracts the unique patient number (UPN), patient’s first and last name, as well as the date of birth. As Excel stores all dates as integer values the function `DateTime.FromOADate` converts the integer value into the appropriate date format. The entire algorithm should exclusively work for patients receiving allo-HSCT. A selection was performed by filtering of table column “KMT_ART”. A number “1” in this column indicates an autologous HSCT. Finally, the method `MySQL.datasetPatient2db()` writes the extracted row as a new dataset into the table `patient` in the database `fac_db`.

¹¹ OLEDB or OLE-DB (object linking and embedding database) is an application programming interface (API) by Microsoft, which allows accessing different data sources.

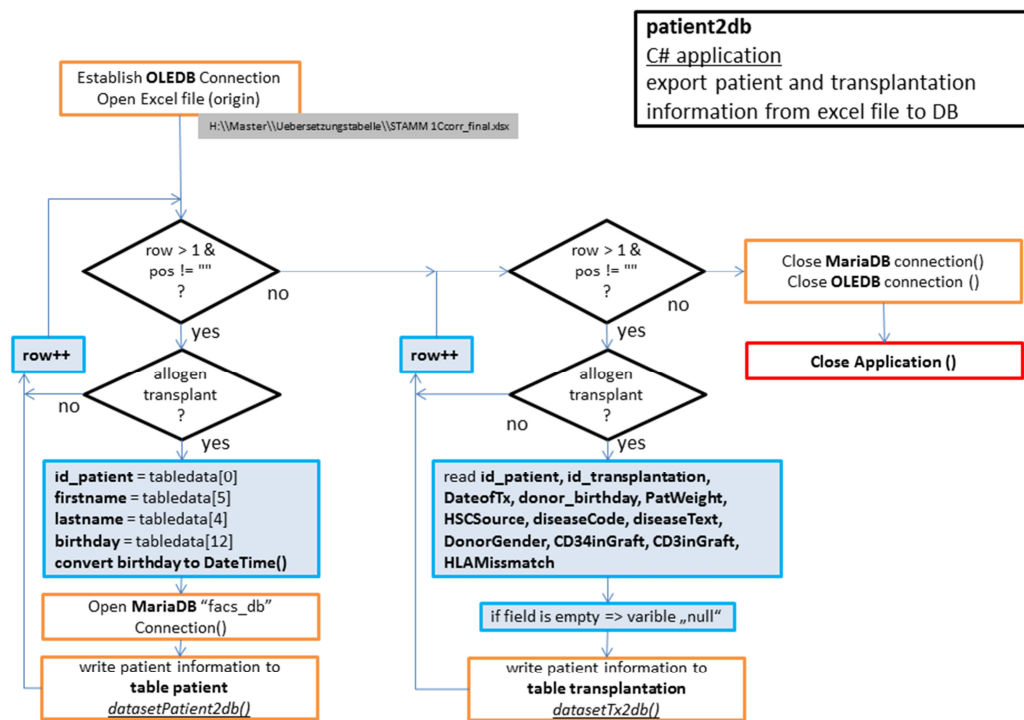


Figure 8: “patient2db” flow chart. The chart describes the program flow of the C# application “patient2db”, which extracts personal identification information of the patient and information related to the HSCT and the donor material

The next section extracts the transplantation relevant data such as UPN, HSCT number, date of HSCT (converted into date format), donor’s date of birth, patient’s weight, disease code, disease text, transplantation material (coded), donor’s gender (coded), donor’s relationship to recipient, count of CD34+ cells in graft, count of CD3+ cells, in graft and the count of HLA mismatch between donor and recipient. During checking of the correct count of CD34+ cells and CD3+ cells new columns were created, where the corrected count of cells was written down. In extraction of the count of CD3+ and CD34+ cells, the algorithm prefers to take the newer column/value. If both fields do not contain a value, the value is set to “NULL”. Finally, the extracted values are sent as a new dataset to the table transplantation in the database facs_db. The method `MySQL.datasetPatient2db()` performs this operation.

Additionally, relevant information and errors were reported to the user in the text console. Since the database uses a dot as a decimal delimiter, the numbers had to be converted to a string value and the dot was replaced by a comma. An empty field was coded with “NULL”.

2.4.5.2 *parse_xls*

The second object-oriented application to extract data is focused on the parsing of the patient's flow cytometry reports (see sections 2.1 and 2.4.3.1). A simplified flow chart about the application can be found in Figure 9. The source of the prepared flow cytometry file (section 2.4.3.1) has to be manually defined at the beginning of the application. Since datasets in the table `analysis` had to be connected to the according patient and according transplantation, the foreign key `id_patient` is queried of the table `transplantation`; therefore, a record of analysis is unequivocally assigned to a unique transplantation and to one unique patient. The files were manually prepared, so that the first line of analysis results start at line 9 in the Excel table. Therefore the algorithm starts extracting data at line 9. To prevent reading of empty lines, the next step of the application is to verify, if the data field for granulocytes and WBC is empty, or if the remark data field contains the note "SORT". If this was false, the row was ignored. The note "SORT" indicates that a chimerism analysis is placed in the current line and that data fields contain a numeric value, followed by the character "%" and an alphabetic character. The value and the letter were split up, additional signs ignored, and both values stored in separate variables. For example, if the data field "remark" contains the note "sort-pcr" and the data field for white blood cells (WBC) contains the content ">99%d", the ">" was ignored, the value "99" stored in `total_wbc` and the letter "d"(donor) as an additional information was stored in `total_wbc_sort`. The variable for analysis method was set to "sort-pcr". This result represents a chimerism analysis, based on PCR, and indicates (more than) 99% of the detected donor cells in the WBC cohort. The analysis method of sorted cells can also be FISH. No additional note stands for a quantitative flow cytometry analysis. The value for the analysis method is stored in the variable `analysemethode`. The remark field can also indicate other specimens like "KM" for bone marrow or "liquor" (details see Table 3). This information is stored in variable `material`. To sum up, the column `remark` indicates the analysis method and the specimen.

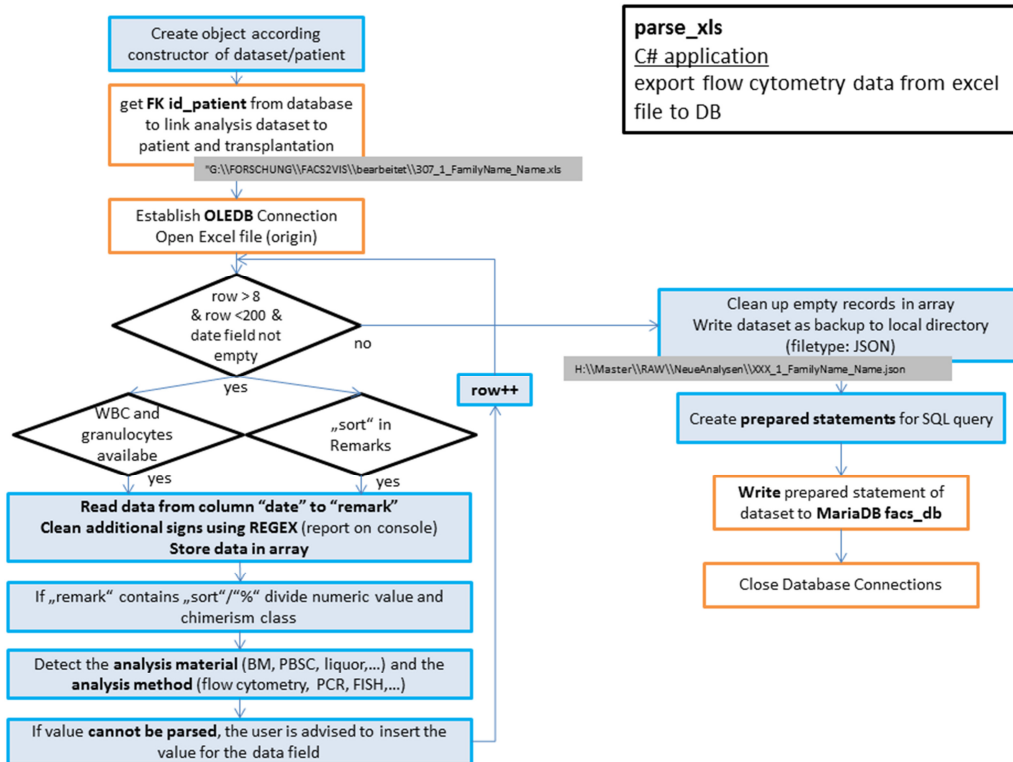


Figure 9: Simplified “parse_xls” flow chart. The chart describes the program flow of the C# application “parse_xls”, which extracts flow cytometry data of Excel records, filters them and stores the relevant data in a structured manner to the database “facs_db”

If a data field in the table is empty or cannot be assigned, the user is asked by a prompt to enter a value (default: “NULL”). Additional characters like “>,(,?,<” were ignored by regular expressions(regex)¹². To develop the adequate filters, the online regex tester and debugger on <https://regex101.com> was used (see screenshot of webpage in Figure 10). To filter numbers with unlimited pre-decimal count of digits and 0-5 post-decimal positions, the regex string `[0-9]+\.,*[0-9][14]` was implemented, being aware that low numeric counts (e.g. “(<1?!”) are coded with “1” and “lager-than-counts” (e.g. >95%) coded with “95” (example see Figure 11).

¹² Regular expression or regex is a formal description language to define search pattern.

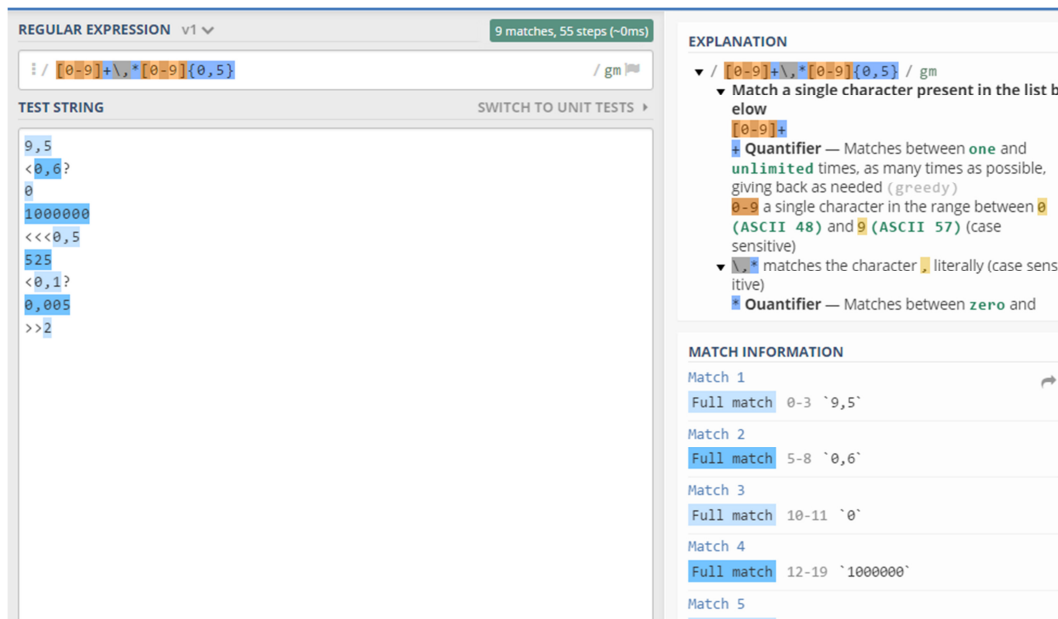


Figure 10: Screenshot of regex101.com online regex tester and debugger. This tool was used to develop regex strings for pattern finding of data input samples. In this example additional non-numeric characters are not adopted in the match results.

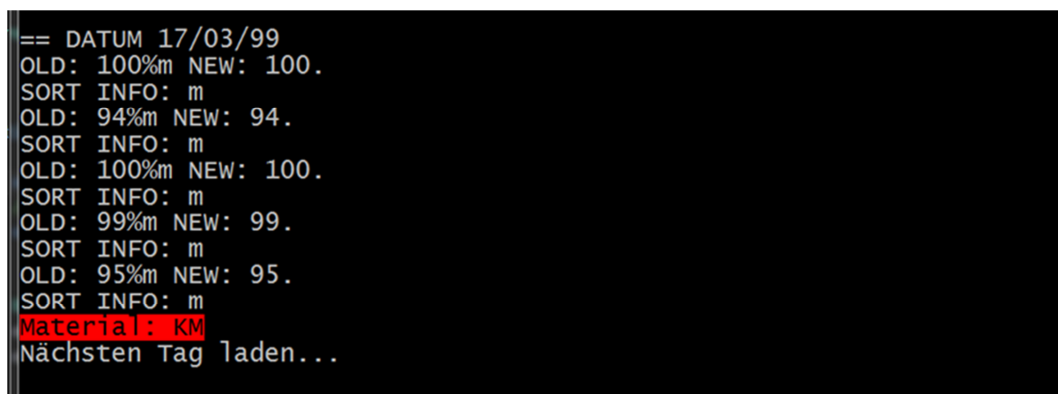


Figure 11: parse_xls console prompt. The program reports to the user if original data was changed. In this case the dataset contains chimerism results, therefore the original information is split up into a numeric variable and an additional “sort info”.

2.4.6 Website development and setup

A web-based graphical user interface (GUI) was developed (1) to filter and visualize database datasets, (2) to report and visualize comparison of groups and (3) to have a mask for input of analysis results (see Figure 12). The GUI and functionality was developed using Hypertext Markup Language (HTML), Java Script (js), Cascading Style Sheets (CSS), and Hypertext Preprocessor (PHP). With PHP the server-side connection to the database was performed, where other technologies were mainly used for the data processing and visualization on

HTML pages. As an additional support the following two external libraries were used:

- [jQuery](#) [32] is a free on js-based library which allows Document-Object-Model manipulation (manipulation of HTML page content). It facilitates connection to server-based applications or pages by the technology of Asynchronous JavaScript and XML (AJAX). [Jquerymobile](#) provides additional features for interactivity and appearance on mobile devices with touch screens.
- [C3.js](#) (details see section 1.1.5) is a js library based on D3.js. It provides a framework and tools for data driven documents. In this project the tool for visualization and interactivity of time series is used.

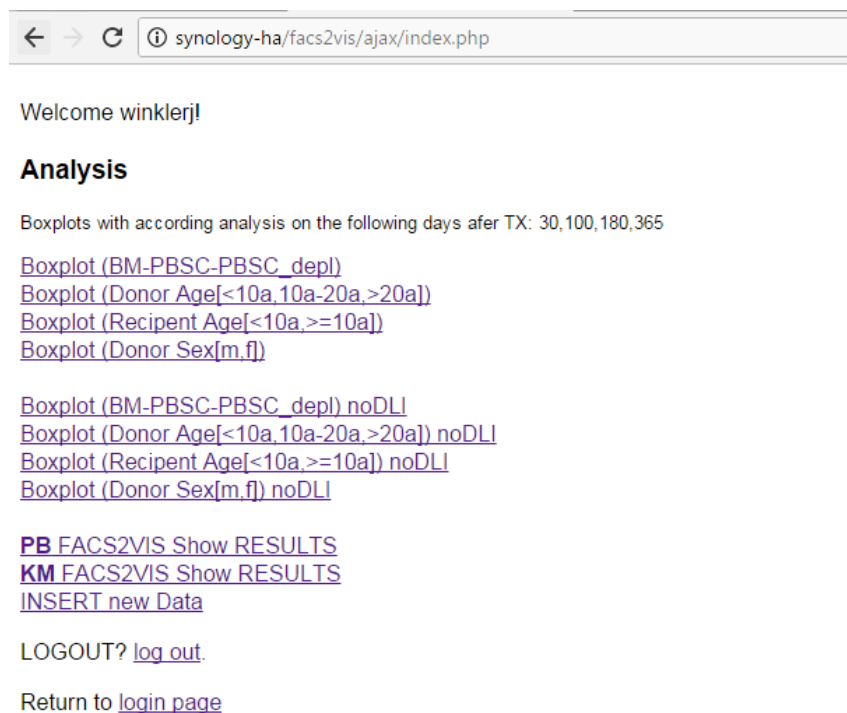


Figure 12: Main menu (landing page) of work related to the master thesis after successful login procedure. The links in the first two blocks guide the user to boxplots and statistical analysis. “PB FACS2VIS Show RESULTS” shows datasets of PBSC flow cytometry results. “KM FACS2VIS Show RESULTS” shows datasets of BM flow cytometry results. “INSERT new Data” opens a mask to insert new measurement results.

The entire development took place in the graphical development environment of “WebStorm 2016.3.4” and first tested on the local computer with XAMPP and since 13/02/2017 on the CCRI server `synology-ha` with PHP Version 5.6.30.

To protect the webpages from unauthorized access, all pages start with an execution of a function, which proofs, if a user is logged in to see the page. This code is executed at first at every page: `<?php if (login_check($mysqli)`

== true) : ?> If no user is logged in the user cannot access the page content, but gets a message to continue with the login procedure. The code and implementation for the login procedure was adopted from [30].

2.4.6.1 Technologies of data transport

The web application requires access to the database `facts_db` for (1) verification of page access authentication, (2) obtainment of selected information of the database and (3) to write new datasets to the database. Therefore the application is designed to have php pages with embedded html and js code presented to the user during page access and additional server-based php pages, which can provide additional information including database access. The communication between pages on the client's side and server-side applications is implemented by the technology of AJAX. For example, if the user requires additional information about an analysis, the unique patient identification number is asynchronously hand over to the server-application. The server application processes the request and delivers the result to the client. In Listing 1 a code snippet of an AJAX call is displayed. `$("#selectPatient").load('getMenue.php?data='+tmp)` loads an dropdown menu. The entries of this menu are generated in `getMenue.php`, according to the submitted settings of the json string `tmp`.

Listing 1: AJAX method to load returned data into patient dropdown menu. The preselected variables are hand-over in a json string to the file `getMenue.php`

```
var jsonObject =
{
    hla_min: hla_min,
    hla_max: hla_max,
    age_min: age_min,
    age_max: age_max,
    sex_male: sex_male,
    sex_female: sex_female,
    graft_bm: graft_bm,
    graft_pbsc: graft_pbsc,
    cd34_min: cd34_min,
    cd34_max: cd34_max,
    cd3_min: cd3_min,
    cd3_max: cd3_max,
    cd3null: cd3null,
    cd34null: cd34null
}
var tmp = JSON.stringify(jsonObject);
console.log(tmp);
$("#selectPatient").load('getMenue.php?data='+tmp)
```

2.4.6.2 Filtering and visualization

The page `facts2vis.php` and `facts2vis_KM.php` are basically equal. The only difference is that `facts2vis.php` refers to the results of flow cytometry-based analysis of peripheral blood samples, whereas `facts2vis_KM.php` display results of bone marrow samples. On both pages, preselection for the displayed

patients in the dropdown menu below “Please select Patient” can be made with (range) sliders or buttons. With range sliders, a minimum and a maximum of a range can be defined.

At “Human Leukocyte Antigen (HLA) Mismatch [out of 10]” the maximal and minimal count of HLA mismatch between donor and recipient can be defined via a range slider. The buttons “Donor’s sex” can restrict the displayed patients according to male and female donors. The buttons in section “Origin of GRAFT” allows filtering of patients according to donor’s material (BM or PBSC). The range sliders $\text{CD34/kgBM}[10^6/\text{kg}]$ and $\text{CD3/kgBM}[10^6/\text{kg}]$ define the minimal and maximal value of stem cells and T-cells in donor’s graft. If the following checkboxes “Show only $\text{CD3/kgBW} = \text{null (unknown)}$ ” respectively “Show only $\text{CD34/kgBW} = \text{null (unknown)}$ ” are checked, only patients where information about $\text{CD34}^+ / \text{CD3}^+$ cells within the graft is missing, are displayed (see Figure 13).

Each time a filter is set, the function `showSelectionMenu()` is called. This javascript function collects the adjusted settings, puts them into a JSON string and calls the `getMenu.php` (see Listing 1) to write entries of the drop-down menu. These entries consist of the following format: “family Name, first name, (*date of birth), date of transplantation description of disease” (e.g. “Hans Mustermann (*1999-12-30) 2000-01-17 SCID”).

“ CD4N TH ” defines the threshold for naïve CD4^+ T-cells, “ CD 19 TH ” the threshold for CD19^+ B-cells. The result of these two settings corresponds to the first day of analysis after transplantation, when the naïve CD4^+ T-cell count / CD19^+ B-cell count is higher than the adjusted threshold.

Figure 13: Screenshot of facs2vis.php

The function `showUser(str)` is executed, when a patient in the list is selected. It stores local selection settings, hands over the ID of the selected patient and the threshold settings and loads results of `getPatData.php`. The algorithm in `getPatData.php` queries the database according to the parameter passed. The extracted datasets, including date, count of sell subtypes, and remarks are displayed in a table on the webpage and loaded into C3.js time series graph. Furthermore results are the two dates, where cell count of naïve CD4+ T-cells / CD19 B-cells are higher than their threshold. If the threshold is never reached, the returned variable is “never”.

The C3 timeseries line chart is bound to `<div id="chart"></div>` at the bottom of `facs2vis.php` and `facs2vis_KM.php` and as an execution of function `showUser(str)` the page scrolls to the top of this chart.

Chart

The line chart with timeseries data was chosen from <http://c3js.org/examples.html> and adapted to the following requirements. Up to 17 lines are displayed in one

graph. To distinguish the categories, different colors are used according to the guidelines at <http://colorbrewer2.org> [33], but, because to the count of lines the color selection of these qualitative (no ranking) dataset is limited. It was not possible to get a colorblind-safe, printer-friendly and photocopy safe color selection. Furthermore, the number of data classes at colorbrewer is limited to 12. Finally the suggested colors of the c3 line chart were supplemented with colors of the colorbrewer suggestion. For a detailed list see Table 8.

Table 8: C3 time series graph color selection

Class name	Color (HEX)	Name of cell type
■ anl_WBC	#000000	White blood cells
■ anl_CD3	#1F77B4	CD3+, T-Lymphocytes
■ anl_CD38	#FF7F0E	CD38+, activated CD3
■ anl_CD4	#2CA02C	CD4+, T-helper cells
■ anl_CD4_naiv	#D62728	naïve CD4+
■ anl_CD8	#A37DC6	CD8+, cytotoxic T-lymphocytes
■ anl_CD8_naiv	#8C564B	naïve CD8+
■ anl_NK	#E377C2	natural killer cells
■ anl_NKT	#7F7F7F	natural killer T-lymphocytes
■ anl_Mono	#BCBD22	monocytes
■ anl_Granulo	#17BECF	granulocytes
■ anl_CD34	#52527A	CD34+, stem cells
■ anl_B_precursor	#B3B3CC	B-cells precursor
■ anl_CD19	#FF00FF	CD19+ B-cells
■ anl_Eos	#99CC00	eosinophil granulocytes
■ anl_gammadelta	#E6B3CC	γδ T-lymphocytes
■ anl_normo	#999900	normocytes (erythrocyte precursors)

All data shown in the graph is based on the unit cells/μl. anl_CD38, anl_CD4_naiv, anl_CD8_naiv, anl_B_precursor, anl_gammadelta and anl_normo are stored as relative values, therefore the absolute count for each cell line has to be calculated before handing-over to the graph. This was performed in a java script and results had to be additional rounded to the second decimal place.

Furthermore the graph supports interactivity (see Figure 14). The displayed lines can be highlighted by a mouse-over at the corresponding legend; the graph can be zoomed in by movement of the mouse wheel and shifted by clicking. An additional subgraph shows a total overview of the whole measurement time period and the selection of the displayed range in the main graph. By clicking on the cell name in the legend the corresponding line can be toggled (visibility). Furthermore, scaling on the x-axis adapts automatically according the maximum value in the displayed range of the graph. A tool tip¹³ is shown when the mouse cursor is moved over a date of measurement. This graph with its interactivity shall facilitates the finding of patterns within the process of cell engraftment.

¹³ Tool tip in this setting: an additional small table with current count of visible cell lines

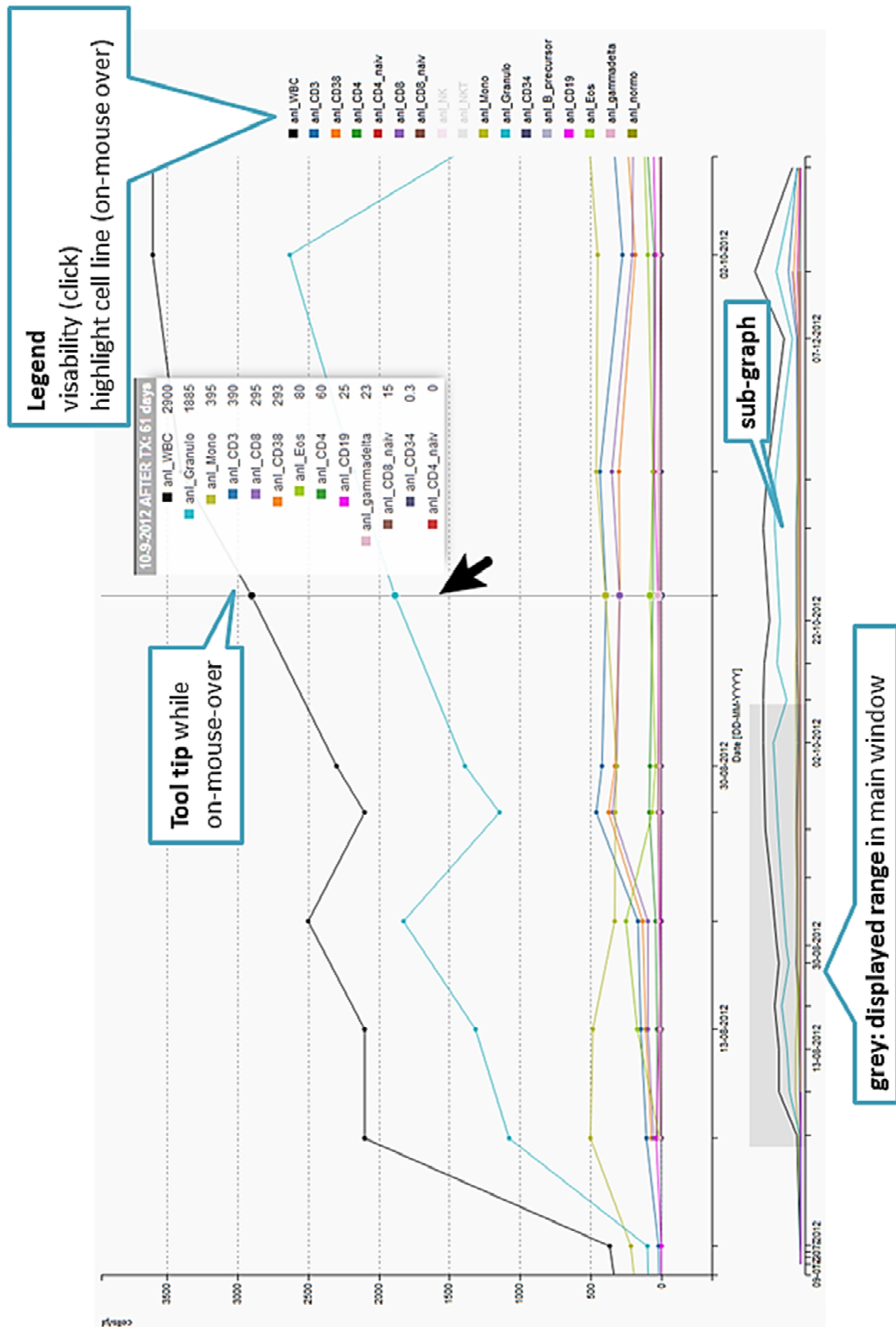


Figure 14: flow cytometry results visualized in an interactive C3 timeseries chart.

2.4.6.3 *Input screen for new data*

The page `insertPatData.php` was developed to insert new data into the database. First, the according transplantation and patient has to be selected. Then, new data can be inserted into a web form. The analysis method and the specimen must be defined. When sending the form, a validity check is performed and the new dataset is sent to the database.

2.4.6.4 *Development of web-based statistical analysis*

Furthermore, a web tool was developed, that shows results of the descriptive statistic and boxplots on the website for group comparison of naïve CD4+ T-cells counts measured via flow cytometry. The code for this `d3.js` boxplots was adopted from [34] and manipulated to the needs of this scope. The design of theses box-plots (or box-whisker-plots) is according to the guidelines of John Wilder Tukey [35]. The lower end of the boxes defines the value for the first quartile, meaning that 25% of all measurements are below this line. The horizontal line inside the box defines the median value (Q2), the upper line of the box the third quartile. Therefore 50% of all measurements are inside the range of the box. The whiskers define 1.5 times the interquartile range and are the limit for outliers. Values outside this range are defined as outliers and marked with a small circle. Numeric values for upper and lower whisker, Q1, Q2 (median), Q3 are noted beside the box or whisker. Additionally the following values are presented in a table: Number of patients (n), maximal value (Max), minimal value (Min), arithmetic mean (Mean), standard deviation (σ), variance, standard error of the mean (SEM), and the kurtosis of the deviation. The heading of each column contains a link where the underlying data can be downloaded as semicolon-separated-file for further statistical processing.

The count of naïve CD4+ T-cells on day 30(+/-5), 100(+/-10), 180(+/-20) and 365(+/-20) after HSCT in different comparison groups was examined. For each patient, only one value at a single time point of measurement was accepted and, concurrently 'NULL' values ignored. So, if more measurement results appear in the range of time points for measurements, the mean value was calculated and processed. The calculation of the mean value took place on the server side by an SQL statement.

2.4.7 **Material and methods for statistical analysis¹⁴**

This section describes the material and methods for answering the following research questions:

¹⁴ **Important note:** The basic explorative statistical analysis in this paper shows an example for data analysis, based on the novel database. To achieve clinical relevance, additional factors for statistical analysis must be considered!

- RQ1: Do graft source, age of donor and age of recipient influence the amount of naïve CD4 T cells on days 30,100,180,365 after allo-HSCT?
 - RQ1.1 Statistical Analysis: Does the graft source (BM, PBSC, CD3 depleted PBSC) influence the engraftment process (absolute cell count) of naïve CD4+ T-cells after allo-HSCT?
 - RQ1.2 Statistical Analysis: Does the engraftment process (absolute count) of naïve CD4+ T-cells correlate with the age of the donor?
 - RQ1.3. Statistical Analysis: Does the engraftment process (absolute count) of naïve CD4+ T-cells correlate with the age of the recipient?

2.4.7.1 Including and excluding criteria

Since 22/11/2000, naïve CD4+ T-cells are detected by the surface marker CD45+/CD45RA+/CD4+/CD62L+. Previously, this cell population was defined by CD45+/CD45RA+/CD45RO-/CD4+, which was stated as imprecise by the biologist Gerhard Fritsch, PhD. However, the methods for detection of naïve CD4+ T-cells with CD45+/CD45RA+/CD4+/CD62L+ is equal to current methode using the following gating strategy: CD45+/CD45RA+/CD4+/CD27+/CD31+ (since 15/07/2013, see Appendix C). Therefore, all flow cytometry analyses for naïve CD4+ T-cells of patients after HSCT, receiving BM or PBSC were included in this statistical analysis, beginning from 22/11/2000. The last included dataset is from 5/12/2016.

Of note, all patients receiving more than one HSCT, CD34+ boost or donor lymphocyte infusion during engraftment were excluded.

2.4.7.2 Definition of comparison groups

The count of naïve CD4+ T-cells on day 30(+/-5), 100(+/-10), 180(+/-20), 365(+/-20) after HSCT on the following group characteristics was examined (overview in Figure 15):

- graft source (BM, PBSC, CD3 depleted PBSC¹⁵)
- donor's age at HSCT (<10a, 10a-20a, >20a)
- recipient's age at HSCT (<10a, ≥10a)

¹⁵ In this content "CD3 depleted PBSC" contain less than $1 \cdot 10^6$ CD3+ cells









graft source	age donor	recipient age
 BM	 <10a	 <10a
 PBSC	 10a-20a	 >10a
 PBSC (CD3 depl)	 >20a	

Figure 15: Comparison groups for statistical analysis

To provide a better basis for comparison the measurement times and the division of group at the age of 10 years was applied according to topic-relevant literature.

2.4.7.3 Statistical test methods and materials

The origin of the data to be analyzed was the export of the semicolon-separated-files (see section 2.4.6.4). These files were exported automatically to the root directory of the website when accessing the website. Then these files were imported to R to be processed.

The following analysis was performed in the open source program R version 3.4.0: (1) statistical significance test based on rank sum tests, such as Kruskal-Wallis rank sum test ($\alpha = 0.05$) [36, 37] or Wilcoxon rank sum test [38-40] were used, since independency of comparison groups exists, but no normal distribution of data was assured. (2) Spearman rank correlation tests were used for monotonic relationship between count of naïve CD4+ T-cells and donor/recipient age [41, 42]. The significance level was set to 0.05. All tests were performed on the days (1) 30(+/-5), (2) 100 (+/-10), (3) 180 (+/-20), (4) 365 (+/-20) after HSCT. The strength of correlation is shown in Table 9.

Table 9: definition of strength of correlation
(Spearman's correlation coefficient ρ), adopted from [43]

Spearman's correlation coefficient ρ	strength of correlation
$0,0 \leq \rho \leq 0,2$	"very weak"
$0,2 < \rho \leq 0,4$	"weak"
$0,4 < \rho \leq 0,6$	"moderate"
$0,6 < \rho \leq 0,8$	"strong"
$0,8 < \rho \leq 0,5$	"very strong"

Due to the fact that multiple tests (total 4) were performed on the same dataset a Bonferroni – correction ($p(\text{corr.bonf}) = p \cdot n$) was applied to reduce the accumulation of the alpha-error. Therefore a statement about significant differences can only be made, if $p(\text{corr.bonf}) < 0.05$. In section 3.3.3 the notation " p_{corr} " is used for Bonferroni corrected values.

A pairwise rank sum test on a dataset of three comparison groups was only applied, if the corrected overall probability value was < 0.05 .

2.4.8 Ensuring of data quality

2.4.8.1 Closing data gaps

The file `SAMM1C.xls` contains relevant information about patients and according transplantation setting. It holds columns for the amount of CD3+ and CD34+ cells within the infused graft. The information of these tables was incomplete and different information about CD3+ and CD34+ cells were annotated in the patient's flow cytometry report files and in the `STAMM1C.xls` documentation. Therefore, a matching of the data in the `STAMM1C.xls` with the original graft analysis reports of the CCRI archive was performed by, starting with the transplantation on 16 April, 1996. If the information of CD34+ and CD3+ cells had differed, the data of the original graft analysis report was adopted into a new table named `CD34kg (myeol)` or `TCell1KG neu`. For better comparison of the data with publications the non-lymphoid¹⁶ amount of stem cells was taken for further calculation.

2.4.8.2 Detection of data and date input errors

After final storage to the database, engraftment charts were visually analyzed for extreme outliers. Human's brain and its communication channels can detect patterns, and therefore outliers more, easier, if they are coded as colors or symbols [21]. Therefore, pattern finding is difficult in tables containing only numeric values. Consecutively, graphs were visually checked if outliers of the x-value (cell counts) and of the y-value (date input errors) occurred.

In the historical documentation of flow cytometry reports for data in the column `Granulo`, dots were used to indicate the thousands digit, but also for the decimal place. Therefore the information placed in the database could have been incorrect. Granulocytes form usually the main part of the WBC [4]. But if the line of granulocytes is not placed close to the line of WBC a possible misplaced decimal point could have been occurred. Therefore, a graphical checkup and query of a dataset of the effected entries of the database was performed. For example, a calculation of the sum of the following values of cells were added (CD3, NK, Mono, Granulo, CD34, CD19, Normo) and finally subtracted from the WBC count. If the difference was not between -50 and +50, the dataset was examined in detail.

¹⁶ Gerhard Fritsch. PhD, named the non-lymphoid stem cells „myeloid“, but this cluster also contains multi potent progenitor cells. In this thesis, these cells are indicated as “myeloid stem cells” or “myeloid CD34+ cells”.

Correction of potential data errors in the database took place after consultation with the laboratory staff and comparison with the historical patient's flow cytometry report.

3 Results

In this chapter the results of the different working packages are presented.

Section 3.1 focuses on the results of the data acquisition, data quality assurance and scraping of the historical original flow cytometry records. Additionally, results concerning the transition of data into the database are shown.

In section 3.2 results regarding the web-based visualization tool of flow cytometry analysis and the appropriate finding of data inputs were shown.

Section 3.3 shows the results of the statistical analysis. This section demonstrates the result of the web-based group comparison and the value distribution concerning the engraftment of naïve CD4+ T-cells. Furthermore, the calculation results of p-values for significant differences of these group comparisons are shown. At the end of this chapter the correlation plots between donor age versus naïve CD4+ T-cell count and recipient age versus naïve CD4 T-cell count are presented.

3.1 Data processing

Verification of the correct information about cell analysis of the donor graft, showed, that out of 82/524 data entries for CD3+ T-cells and 290/732 data entries for CD34+ cells were incorrect and had therefore to be rectified. Most of the errors found in the group of CD3+ T-cells were empty fields or copying mistakes. In the group of CD34+ cells most of the wrong entries were based on adoption of the total count of CD34+ cells, instead of the count of myeloid stem cells.

At data scraping was performed on the human-readable documentation file `STAMM1C.xls` to extract information about patient and transplantation. 612 patients' datasets with first name, last name, and date of birth were extracted and successfully stored in the table `patient`. In total, 712 datasets for the table `patient` were extracted. In the following datasets, a value for the mentioned field could not be extracted, due to a missing value in the corresponding data field of the table `STAMM1C`. Therefore, 53 datasets of `kg_kmt`, 127 datasets of `tx_CD34_kg`, 184 datasets of `tx_CD3_kg`, were assigned to the value `'NULL'`. In one case, the value for the column `dateOfTX` could not be

extracted, due a wrong date format. The date was corrected and successfully transmitted to the database.

26492 datasets were extracted from historical Excel files, regarding flow cytometry analysis data and chimerism analysis. These datasets were stored in the database table called `analysis`. This table contains 16744 datasets related to a flow cytometry analysis and 9748 related to chimerism analyses. Stored analyses were taken during the time period, starting from 28/02/1996 until 16/05/2017 (state 22/06/2017). The content of the tables for the disease code `diag_code` and the table `HSC_Source` was transmitted manually from the description of the corresponding columns of the documentation file `STAMM1C.xls`.

Database and its relation between tables

The result of the database design and its relations between tables is shown in the enhanced entity–relationship (EER) model in Figure 16. A unique dataset is related only to one transplantation and one patient (1:n, one-to-many relationship), thus a patient can have more than one related datasets in the table `analysis` and a transplantation can have more than one related datasets in the table `analysis`. The relationship between table `patient` and table `analysis` is 1:n, also for table `patient` and table `transplantation`. Therefore, one `analysis` dataset and exactly one transplantation are related to one patient, but a patient can have many datasets for flow cytometry analysis and many transplantations. Regarding table transplantation, the relation between table transplantation and table `diag_code` and `HSC_Source` is n:1. One dataset in the table transplantation is related exactly to one `diag_code` and to one `HSC_Source`. But the different disease codes in table `diag_code` and stem cell sources in table `HSC_source` can be used more than once.

Therefore, a relational database, based on open source MariaDB, was implemented, which contains the transmitted information of the separate Excel files, containing patient's specific flow cytometry analysis results.

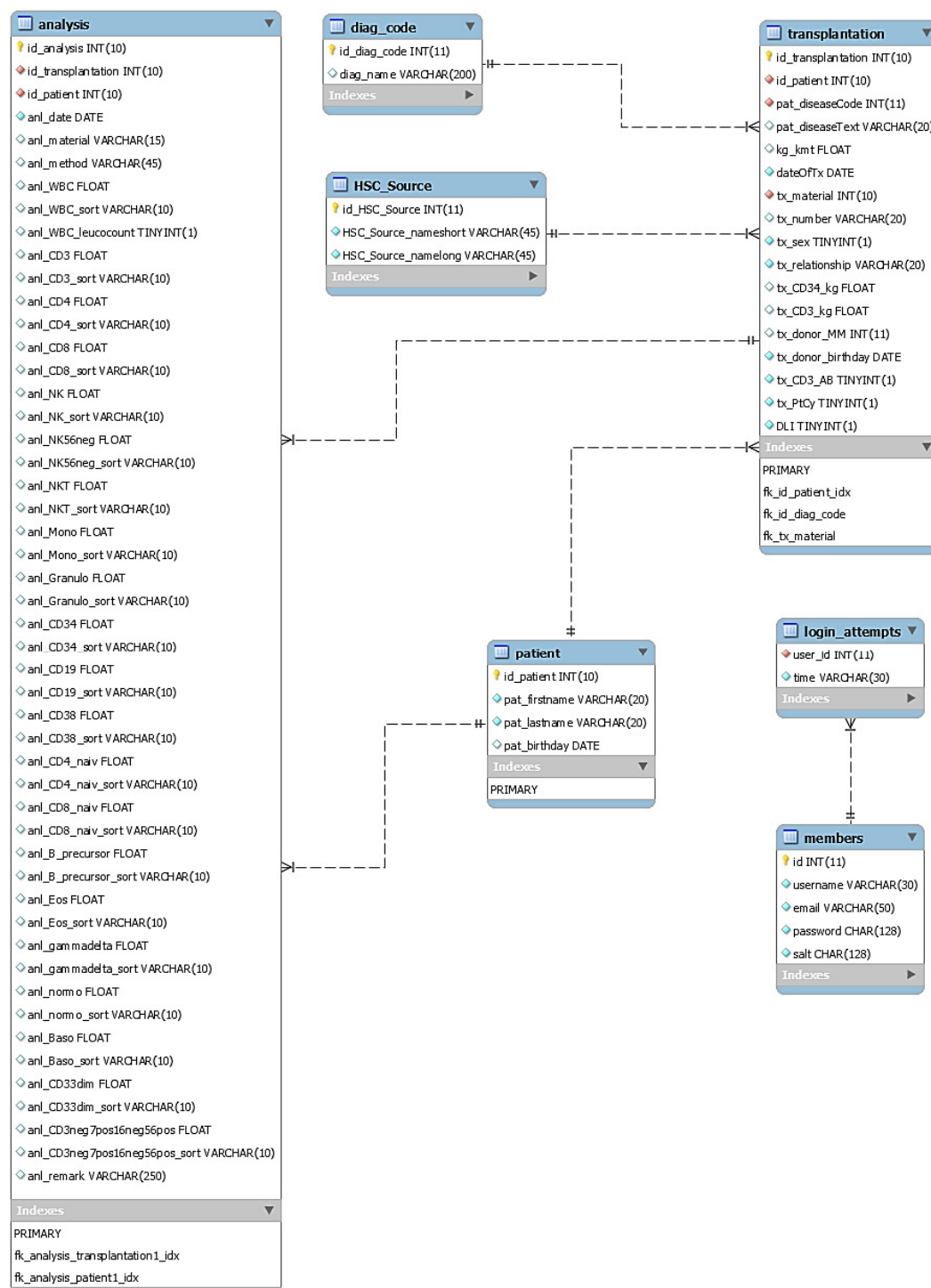


Figure 16: enhanced entity–relationship (EER) model of the database “facs_db”

3.2 Interactive visualization

The interactive web-tool for visualization and insertion of flow cytometry data was successfully implemented. Appropriate methods are described in section 2.4.6.

Both websites (Data visualization page <http://synology->

[ha/facs2vis/ajax/facs2vis.php](http://synology-ha/facs2vis/ajax/facs2vis.php) and webpage for insertion of new data (<http://synology-ha/facs2vis/ajax/insertPatData.php>) were tested in the browsers Internet Explorer 11.0.9600.18697, Mozilla Firefox 54.0 (32-bit) and Google Chrome 58.0.3029.110 (64-bit). No issues in representation and functionality could be found.

Beside visualization, graphical error detection was applied. Detection of the date and data errors are described in the following paragraphs.

Specimen not declared

During visual check of the time lines input errors were identified, because of missing declaration of specimen. Therefore, bone marrow was misinterpreted as peripheral blood. Bone marrow contains usually more cells than peripheral blood. Therefore, this wrong assignment could be identified by spike detection in the line diagram. Usually, bone marrow and peripheral blood is examined at the same day but without declaration of bone marrow specimen. As a result, measurements of different specimens result in an overlay of data points. An example is shown in Figure 17.

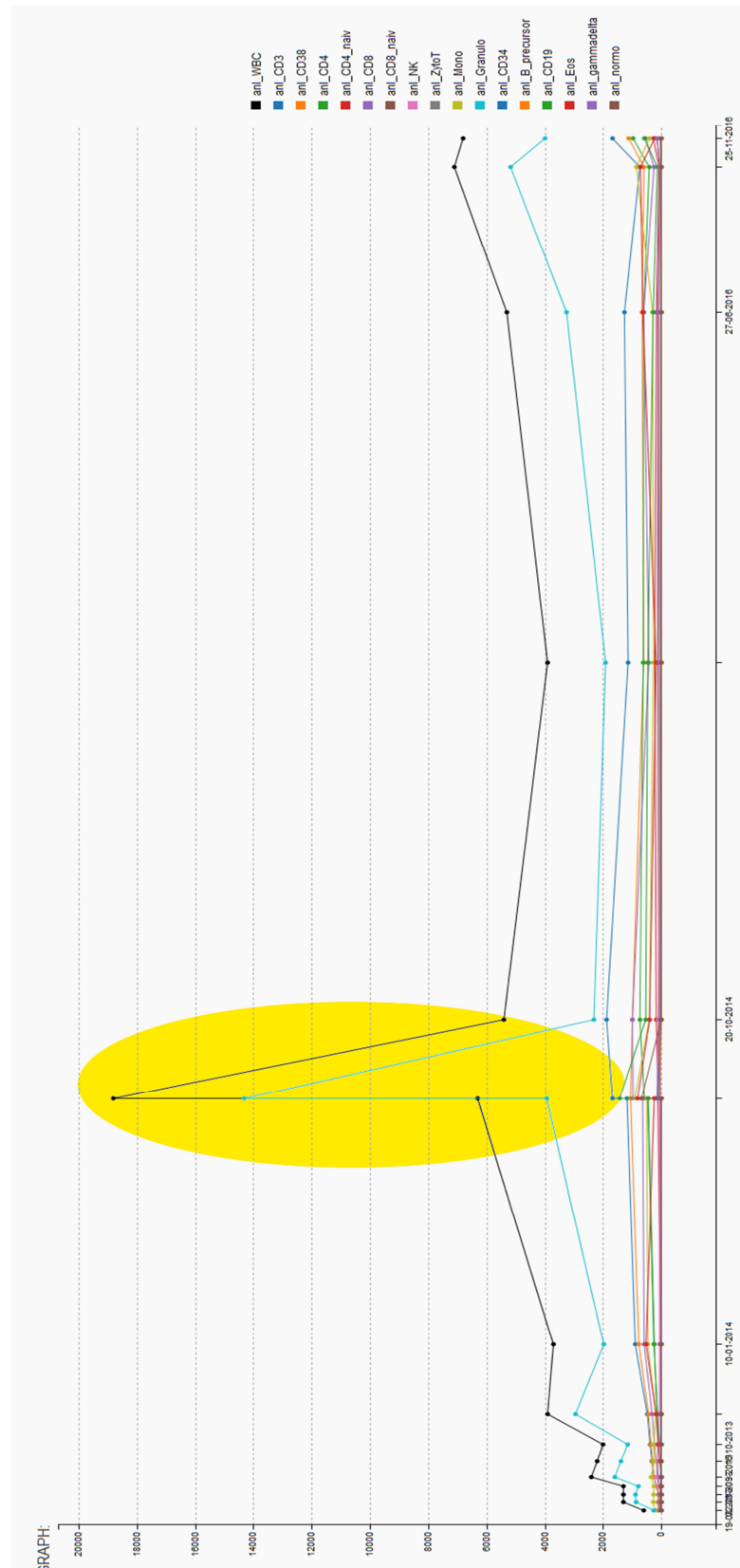


Figure 17: Error detection. Missing declaration of specimen results in a spike / overlay (highlighted in yellow) in the line graph

Date error

During HSCT procedure analysis could also take place a few days before day 0 (day of HSCT), but a big gap between measurement points might include an error, which can be easily detected via visual checking. Therefore, all graphs were analyzed visually to locate and correct input errors. In Figure 19, the big gap between the measurements in the first week after transplantation (points are close in this area) and the first data points (551 days before HSCT), assume a date input error. A comparison with the date stamp in the historical Excel files confirmed a date input error. This error was not detected over years, since the representation of data sets were in table format. The origin of the error is shown in a snippet of the Excel documentation table in *Figure 18*. After date correction, the shape of the line graph assumes a correct progress of cells (see Figure 20).

46	11.03.08	/129	2.900	810
47	31/03/08	/149	5.300	2170
48	31/03/08			
49	30.4.06	179	5.900	725
50	6.5.08	185	3.400	715
51	6.5.08			
52	14.5.08	193	3.800	1120
53	14.5.08			
54	20.5.08	/108	4.600	800

Figure 18: Snippet of the Excel documentation table of flow cytometry results. The date error is highlighted in yellow. The first column shows the dates of analyses, the second the days after HSCT, the third the total count of WBC (cells/ μ l) and the column on the very right the count of CD8+ T-cells/ μ l.

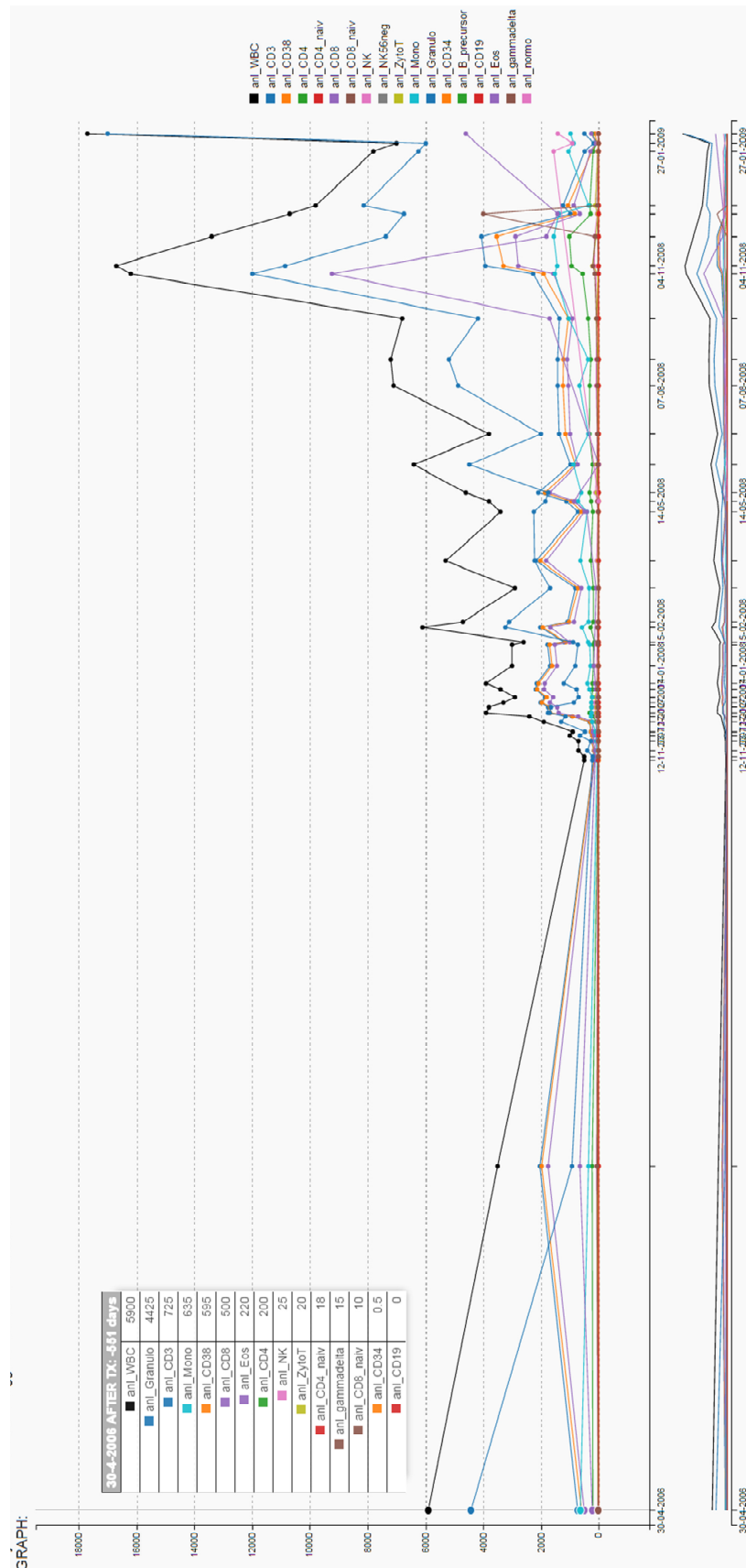


Figure 19: Line chart with date input errors. Two measurements are detected as implausible because of the implausible information: 551 days before HSCT

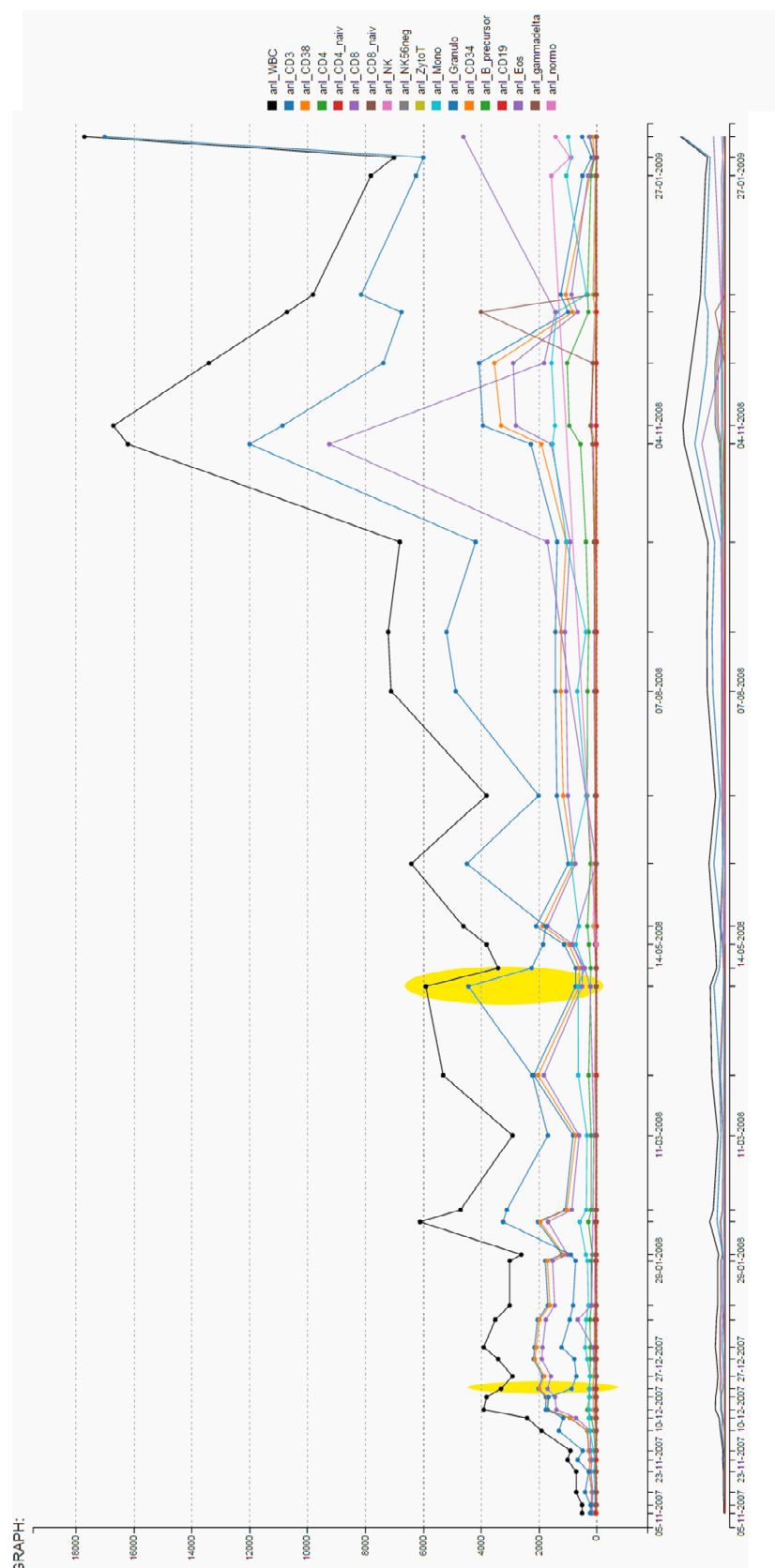


Figure 20: Line chart after correction (highlighted in yellow) of date input errors. Two measurements are detected as implausible because the staining might have been occurred -551 days before transplantation

3.3 Results of statistical analysis

This section describes the results of the statistical analysis. The first part (section 3.3.1) is focused on the web-based descriptive statistics and contains screenshots and snippets of the representation on the website. These screenshots are taken from Mozilla Firefox 54.0 (32-bit).

The second part of this section is based on the exported raw data, provided on the website for descriptive statistics. This raw data is processed in the program R. First, an analysis of normal distribution (section 3.3.2), and then the results of the group comparison tests (section 3.3.3) are shown. In section 0 the results of the correlation tests, regarding donor's age, recipient's age and count of naïve CD4+ T-cells are illustrated.

3.3.1 Web-based descriptive statistics

The following sub sections describe the data distribution of test groups. Every group comparison is done on the days 30(+/-5), 100(+/-10), 180 (+/-20) and 365 (+/-10). The absolute count of naïve CD4+ T-cells [μ l] is shown on the y- axes in all graphs.

3.3.1.1 Graft source

The number of patients measured at day 30 after HSCT vary between the three comparison groups ($n_{BM30} = 194$, $n_{PBSC30} = 43$, $n_{PBSC_CD3_depl30} = 14$). The median count of naïve CD4+ T-cells/ μ l was 0 (range: 0-160, Q1= 0, Q3= 4.75) for BM, 0 (range: 0-142, Q1= 0, Q3= 1) for PBSC, and 0 (range: 0-15, Q1= 0, Q3= 0) for PBSC_CD3_depl. Descriptive analysis of the comparison groups for day 30 (+/- 5) after HSCT (class graft material) is shown in Figure 21.

Day 30 (+/- 5) after TX

	<u>BM</u>	<u>PBSC</u>	<u>PBSC CD3 depl</u>
n	194	43	14
Max	160	142	15
Min	0	0	0
Q1	0	0	0
Median	0	0	0
Q3	4.75	1	0
Mean	5.65	10.70	1.07
σ (std.dev)	15.46	31.93	3.86
variance	239.04	1019.61	14.92
SEM	1.11	4.87	1.03
Kurtosis	55.89	12.60	12.08

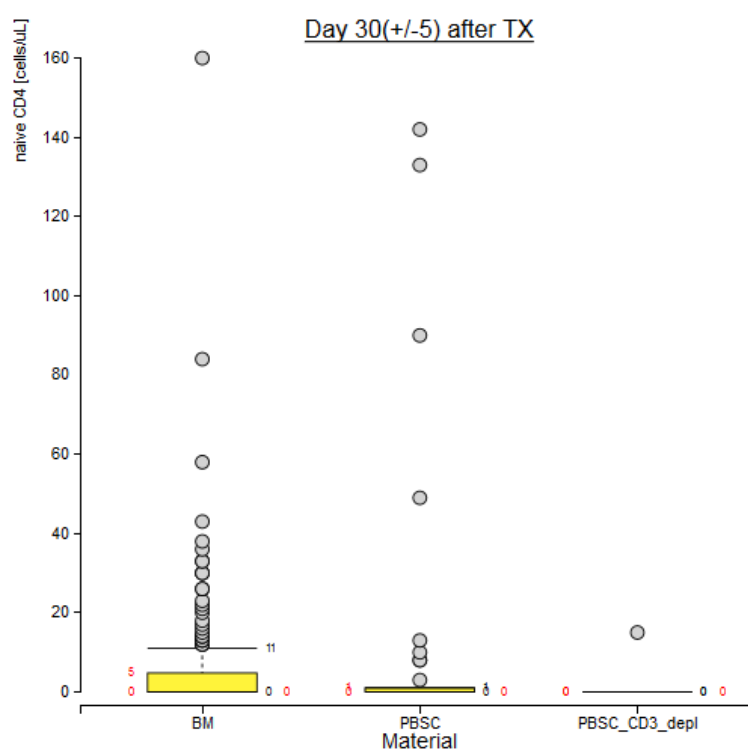


Figure 21: Graft source comparison, 30 (+/- 5) days after HSCT (TX)

The number of patients measured at day 100 after HSCT vary between the three comparison groups ($n_{BM100} = 173$, $n_{PBSC100} = 29$, $n_{PBSC_CD3_depl100} = 17$). The median count of naïve CD4+ T-cells/ μ L was 2 (range: 0-215, Q1= 0, Q3= 8) for BM, 0 (range: 0-143, Q1= 0, Q3= 9) for PBSC, and 0 (range: 0-201, Q1= 0, Q3= 0) for PBSC_CD3_depl. Descriptive analysis of the comparison groups for day 100(+/- 10) after HSCT (class graft material) is shown in Figure 22.

Day 100 (+/- 10) after TX

	BM	PBSC	PBSC CD3 depl
n	173	29	17
Max	215	143	201
Min	0	0	0
Q1	0	0	0
Median	2	0	0
Q3	8	9	0
Mean	9.28	14.93	17.18
σ (std.dev)	22.80	32.03	49.99
variance	519.69	1025.72	2499.44
SEM	1.73	5.95	12.13
Kurtosis	43.07	9.72	10.95

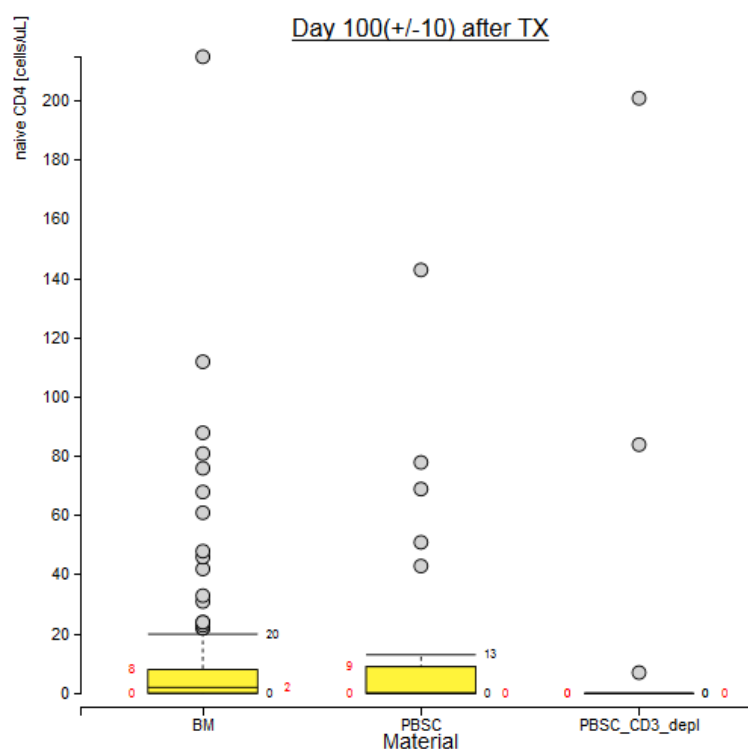


Figure 22: Graft source comparison, 100 (+/-10) days after HSCT (TX)

The number of patients measured at day 180 after HSCT vary between the three comparison groups ($n_{BM180} = 139$, $n_{PBSC180} = 24$, $n_{PBSC_CD3_depl180} = 18$). The median count of naïve CD4+ T-cells/ μ l was 19 (range: 0-1252, Q1= 4, Q3= 74) for BM, 11 (range: 0-251, Q1= 0, Q3= 48.25) for PBSC, and 1.5 (range: 0-1033, Q1= 0, Q3= 20.75) for PBSC_CD3_depl. Descriptive analysis of the comparison groups for day 180(+/-20) after HSCT (class graft material) is shown in Figure 23.

Day 180 (+/- 20) after TX

	BM	PBSC	PBSC CD3 depl
n	139	24	18
Max	1252	251	1033
Min	0	0	0
Q1	4	0	0
Median	19	11	1.5
Q3	74	48.25	20.75
Mean	69.04	43.38	126.17
σ (std.dev)	147.18	67.64	315.79
variance	21662.88	4574.90	99722.47
SEM	12.48	13.81	74.43
Kurtosis	34.68	5.24	6.97

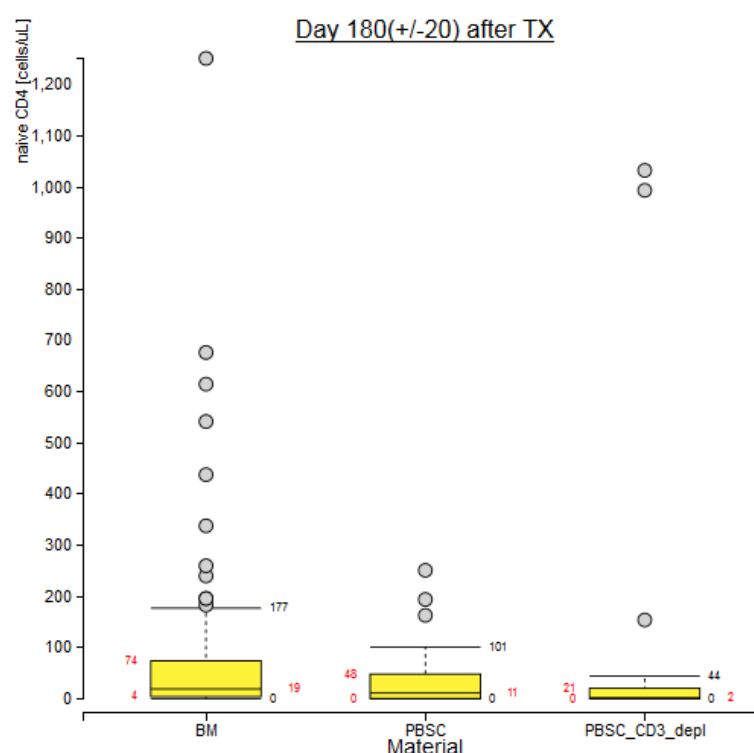


Figure 23: Graft source comparison, 180 (+/-20) days after HSCT (TX)

The number of patients measured at day 365 after HSCT vary between the three comparison groups ($n_{BM365} = 108$, $n_{PBSC365} = 18$, $n_{PBSC_CD3_depl365} = 7$). The median count of naïve CD4+ T-cells/ μ l was 153 (range: 0-996, Q1= 43.75, Q3= 303.5) for BM, 130.5 (range: 0-1455, Q1= 82.5, Q3= 394.25) for PBSC, and 253 (range: 3-1537, Q1= 72.5, Q3= 1184) for PBSC_CD3_depl. Descriptive analysis of the comparison groups for day 365(+/-20) after HSCT (class graft material) is shown in Figure 24.

Day 365 (+/- 20) after TX

	BM	PBSC	PBSC CD3 depl
n	108	18	7
Max	996	1455	1537
Min	0	0	3
Q1	43.75	82.5	72.5
Median	153	130.5	253
Q3	303.5	394.25	1183
Mean	210.00	270.06	614.86
σ (std.dev)	210.64	334.41	612.98
variance	44370.37	111832.83	375739.55
SEM	20.27	78.82	231.68
Kurtosis	5.49	8.90	1.36

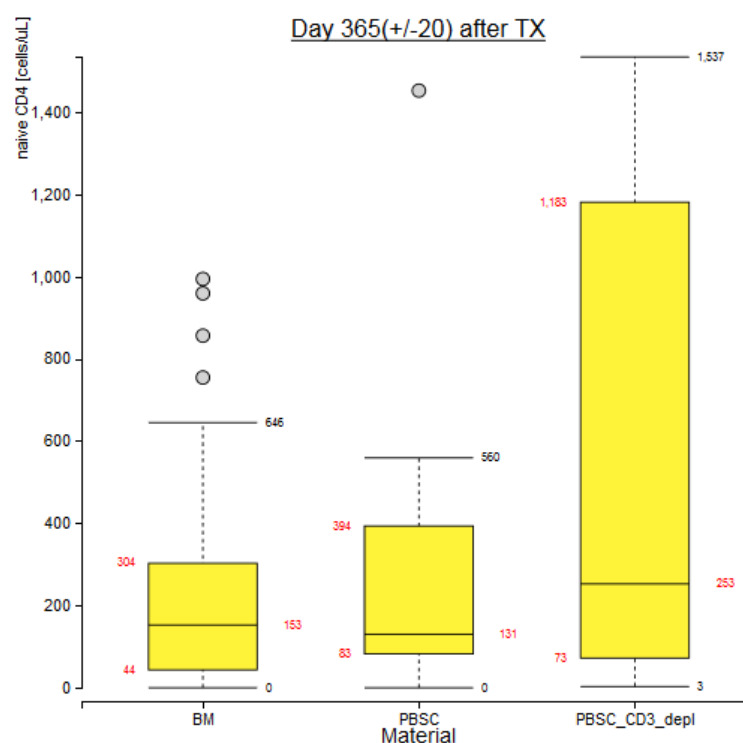


Figure 24: Graft source comparison, 365 (+/-20) days after HSCT (TX)

3.3.1.2 Donor's age

The number of patient's within the three comparison groups – related to 30 days after HSCT - is unequal ($n_{\text{don_younger10a30}} = 39$, $n_{\text{don_10_20a30}} = 40$, $n_{\text{don_older20a30}} = 180$).

The median count of naïve CD4+ T-cells/ μl was 4 (range: 0-58, Q1= 0, Q3= 4) for “Donor's Age <10a”, 5 (range: 0-84, Q1= 0, Q3= 14.5) for “Donor's Age 10a – 20a” and 0 (range: 0-160, Q1= 0, Q3= 0) for “Donor's Age >20a”. Descriptive analysis of the comparison groups for day 30 (+/-5) after HSCT (class donor's age) is shown in Figure 25.

Day 30 (+/- 5) after TX

	Donor's Age <10a	Donor's Age 10a - 20a	Donor's Age >20a
n	39	40	180
Max	58	84	160
Min	0	0	0
Q1	0	0	0
Median	4	5	0
Q3	14.5	14.25	0
Mean	10.59	11.65	4.36
σ (std.dev)	13.88	17.18	20.14
variance	192.75	295.28	405.82
SEM	2.22	2.72	1.50
Kurtosis	5.12	9.50	43.11

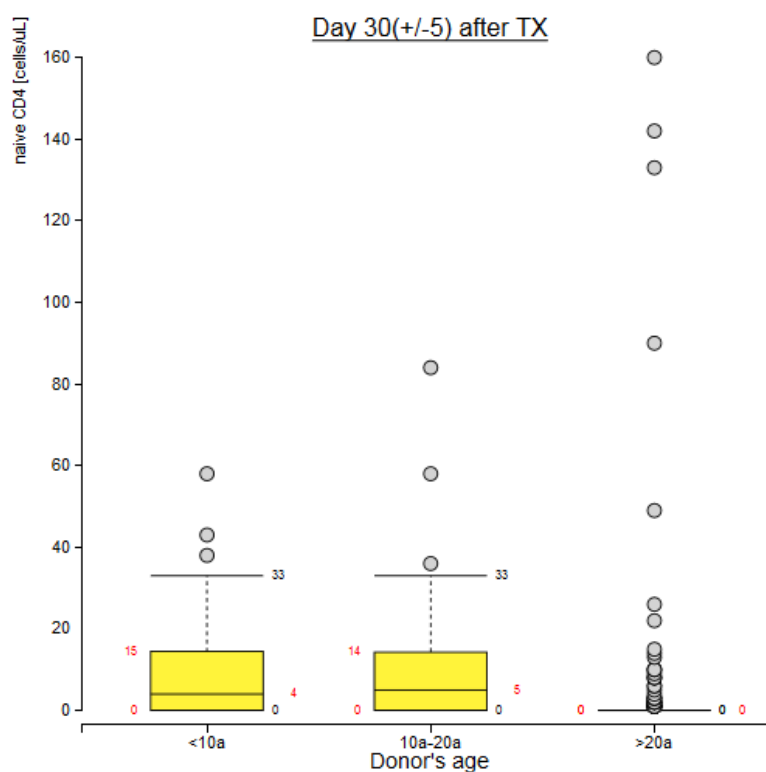


Figure 25: Donor's age comparison, 30 (+/-5) days after HSCT (TX)

The number of patient's within the three comparison groups – related to 100 days after HSCT - is unequal ($n_{\text{don_younger10a100}} = 36$, $n_{\text{don_10_20a100}} = 38$, $n_{\text{don_older20a100}} = 151$). The median count of naïve CD4+ T-cells/ μL was 9 (range: 0-239, Q1= 2, Q3= 19) for “Donor's Age <10a”, 6.5 (range: 0-215, Q1= 2, Q3= 14.5) for “Donor's Age 10a – 20a” and 0 (range: 0-201, Q1= 0, Q3= 3.5) for “Donor's Age >20a”. Descriptive analysis of the comparison groups for day 100 (+/-10) after HSCT (class donor's age) is shown in Figure 26.

Day 100 (+/- 10) after TX

	Donor's Age <10a	Donor's Age 10a - 20a	Donor's Age >20a
n	36	38	151
Max	239	215	201
Min	0	0	0
Q1	2	2	0
Median	9	6.5	0
Q3	19	14.5	3.5
Mean	22.31	17.05	7.65
σ (std.dev)	42.68	36.41	24.76
variance	1821.99	1325.52	612.92
SEM	7.11	5.91	2.01
Kurtosis	19.06	23.26	33.18

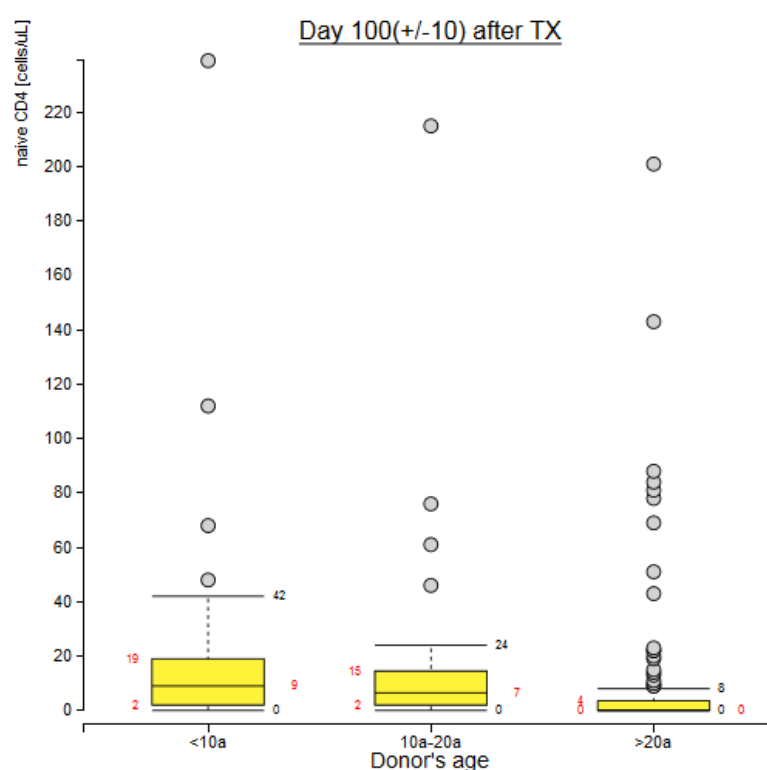


Figure 26: Donor's age comparison, 100 (+/-10) days after HSCT (TX)

The number of patient's within the three comparison groups – related to 180 days after HSCT - is unequal ($n_{\text{don_younger10a180}} = 30$, $n_{\text{don_10_20a180}} = 31$, $n_{\text{don_older20a100}} = 125$). The median count of naïve CD4+ T-cells/ μL was 51.5 (range: 2-1252, Q1= 15.25, Q3= 136) for “Donor's Age <10a”, 24 (range: 0-542, Q1= 11.5, Q3= 65.5) for “Donor's Age 10a – 20a” and 9 (range: 0-1033, Q1= 1, Q3= 50) for “Donor's Age >20a”. Descriptive analysis of the comparison groups for day 180 (+/-20) after HSCT (class donor's age) is shown in Figure 27.

Day 180 (+/- 20) after TX

	Donor's Age <10a	Donor's Age 10a - 20a	Donor's Age >20a
n	30	31	125
Max	1252	542	1033
Min	2	0	0
Q1	15.25	11.5	1
Median	51.5	24	9
Q3	136	65.5	50
Mean	166.30	57.23	59.35
σ (std.dev)	292.98	97.93	148.28
variance	85834.48	9590.43	21986.18
SEM	53.49	17.59	13.26
Kurtosis	9.00	19.50	30.09

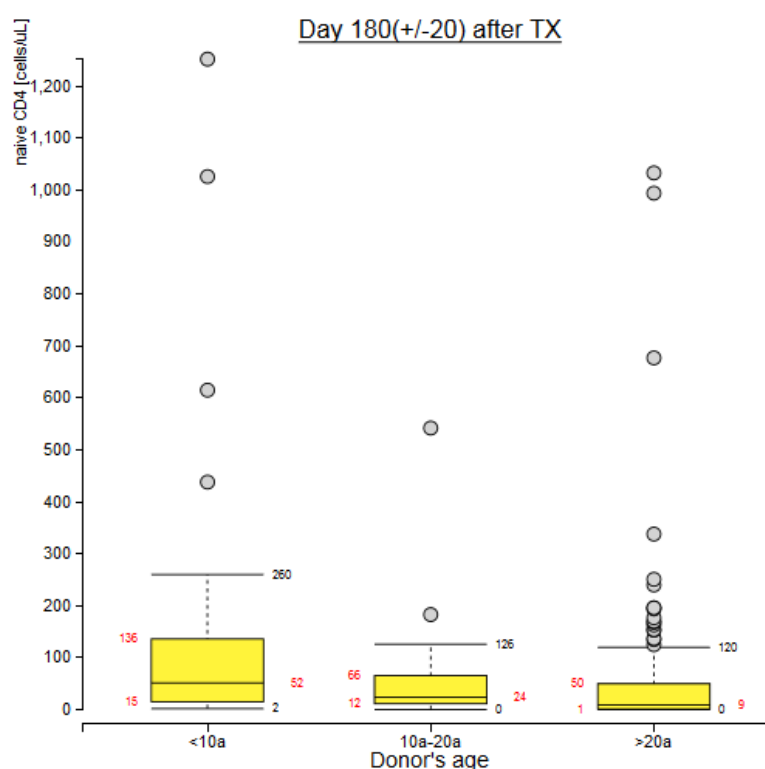


Figure 27: Donor's age comparison, 180 (+/-20) days after HSCT (TX)

The number of patient's within the three comparison groups – related to 365 days after HSCT - is unequal ($n_{\text{don_younger10a365}} = 28$, $n_{\text{don_10_20a365}} = 24$, $n_{\text{don_older20a365}} = 85$). The median count of naïve CD4+ T-cells/ μL was 174.5 (range: 9-996, Q1= 71.25, Q3= 368.5) for “Donor's Age <10a”, 129 (range: 12-511, Q1= 62.5, Q3= 236.75) for “Donor's Age 10a – 20a” and 168 (range: 0-1537, Q1= 32, Q3= 387) for “Donor's Age >20a”. Descriptive analysis of the comparison groups for day 365 (+/-20) after HSCT (class donor's age) is shown in Figure 28.

Day 365 (+/- 20) after TX

	Donor's Age <10a	Donor's Age 10a - 20a	Donor's Age >20a
n	28	24	85
Max	996	511	1537
Min	9	12	0
Q1	71.25	62.5	32
Median	174.5	129	168
Q3	368.5	236.75	387
Mean	247.79	154.54	266.15
σ (std.dev)	221.46	122.43	319.24
variance	49044.03	14988.41	101912.95
SEM	41.85	24.99	34.63
Kurtosis	5.41	3.71	7.68

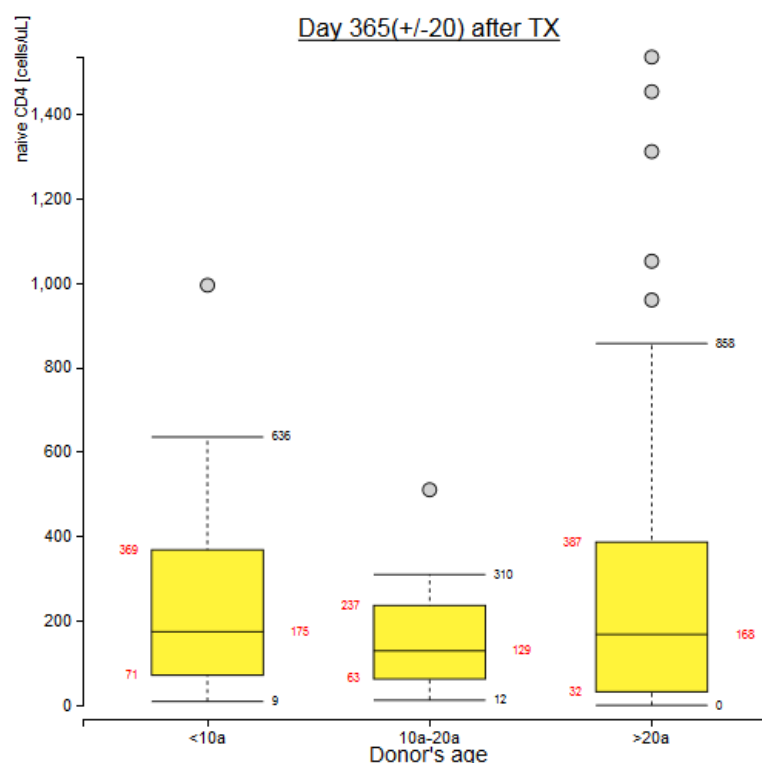


Figure 28: Donor's age comparison, 365 (+/-20) days after HSCT (TX)

3.3.1.3 Recipient's age

The number of patients within the two comparison groups at 30 days after HSCT, is nearly equal ($n_{\text{rec_younger10a30}} = 135$, $n_{\text{rec_older10a30}} = 124$). The median count of naïve CD4+ T-cells/ μL was 0 (range: 0-160, Q1= 0, Q3= 2) for “Recipient's Age <10a” and 0 (range: 0-84, Q1= 0, Q3= 5) for “Recipient's Age $\geq 10\text{a}$ ”. Descriptive analysis of the comparison groups for day 30(± 5) after HSCT (class recipient's age) is shown in Figure 29.

Day 30 (± 5) after TX

	Recipient's Age <10a	Recipients's Age $\geq 10\text{a}$
n	135	124
Max	160	84
Min	0	0
Q1	0	0
Median	0	0
Q3	2	5
Mean	7.64	5.09
σ (std.dev)	23.92	11.79
variance	572.17	139.11
SEM	2.06	1.06
Kurtosis	26.43	21.78

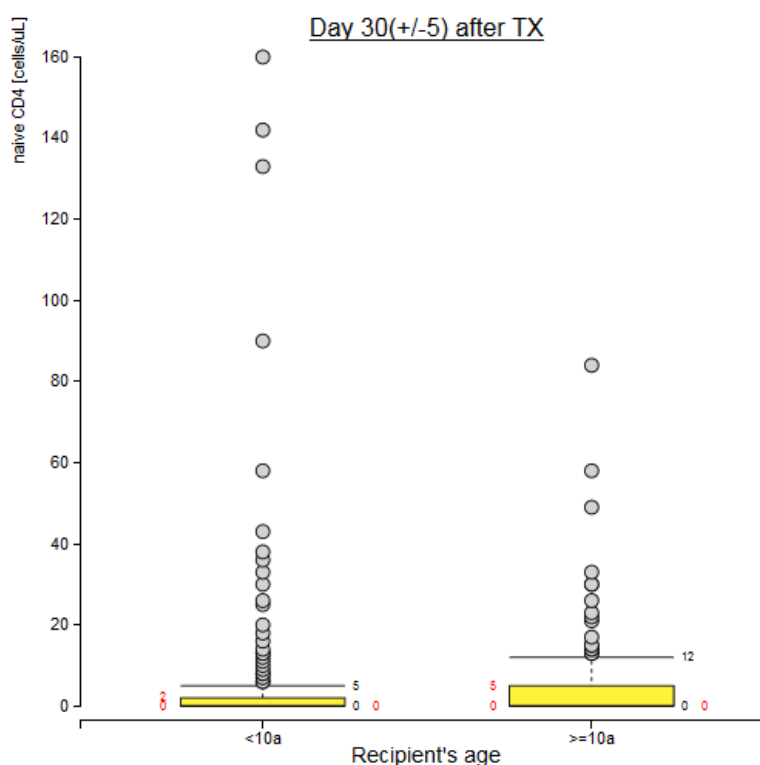


Figure 29: Recipient's age comparison, 30 (± 5) days after HSCT (TX)

The number of patients within the two comparison groups at 100 days after HSCT, is nearly equal ($n_{\text{rec_younger10a100}} = 116$, $n_{\text{rec_older10a100}} = 109$). The median count of naïve CD4+ T-cells/ μl was 2 (range: 0-239, Q1= 0, Q3= 13) for “Recipient’s Age <10a” and 0 (range: 0-84, Q1= 0, Q3= 6) for “Recipient’s Age $\geq 10\text{a}$ ”. Descriptive analysis of the comparison groups for day 100(± 10) after HSCT (class recipient’s age) is shown in Figure 30.

Day 100 (± 10) after TX

	Recipient's Age <10a	Recipients's Age $\geq 10\text{a}$
n	116	109
Max	239	84
Min	0	0
Q1	0	0
Median	2	0
Q3	13	6
Mean	17.00	5.82
σ (std.dev)	40.41	13.43
variance	1633.02	180.43
SEM	3.75	1.29
Kurtosis	17.85	21.46

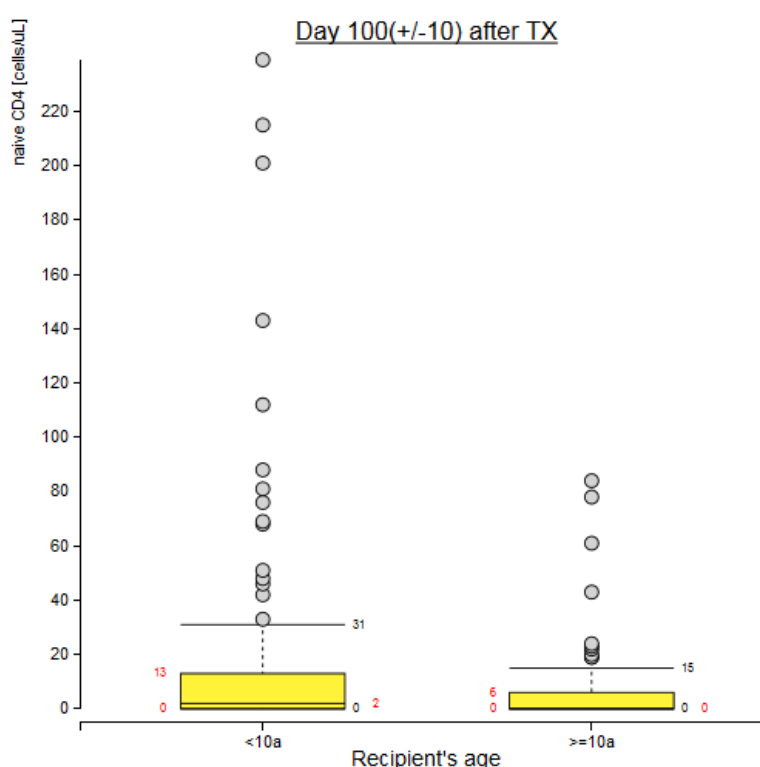


Figure 30: Recipient’s age comparison, 100 (± 10) days after HSCT (TX)

The number of patients within the two comparison groups at 180 days after HSCT, is nearly equal ($n_{\text{rec_younger10a180}} = 94$, $n_{\text{rec_older10a180}} = 92$). The median count of naïve CD4+ T-cells/ μl was 46.5 (range: 0-1252, $Q1 = 7$, $Q3 = 137.75$) for “Recipient’s Age <10a” and 10 (range: 0-170, $Q1 = 1$, $Q3 = 26.5$) for “Recipient’s Age $\geq 10a$ ”. Descriptive analysis of the comparison groups for day 180(+/-20) after HSCT (class recipient’s age) is shown in Figure 31.

Day 180 (+/- 20) after TX

	Recipient's Age <10a	Recipients's Age $\geq 10a$
n	94	92
Max	1252	170
Min	0	0
Q1	7	1
Median	46.5	10
Q3	137.75	26.5
Mean	130.61	20.71
σ (std.dev)	236.71	29.86
variance	56033.79	891.90
SEM	24.42	3.11
Kurtosis	12.33	9.85

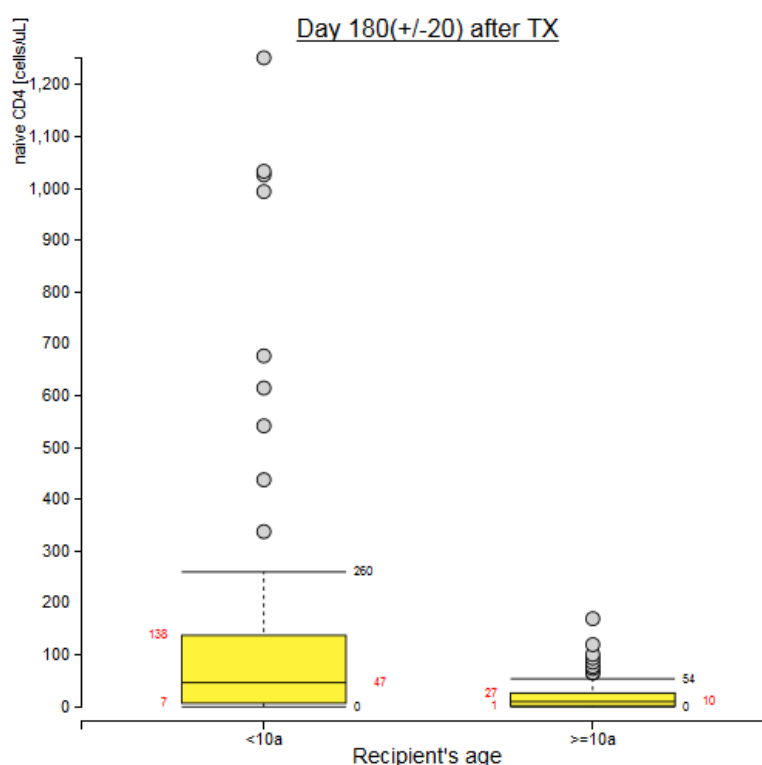


Figure 31: Recipient’s age comparison, 180 (+/-20) days after HSCT (TX)

The number of patients within the two comparison groups at 365 days after HSCT, is unequal ($n_{\text{rec_younger10a365}} = 79$, $n_{\text{rec_older10a365}} = 58$). The median count of naïve CD4+ T-cells/ μL was 302 (range: 1-1537, Q1= 153, Q3= 456) for “Recipient’s Age <10a” and 63 (range: 0-305, Q1= 19.5, Q3= 139.5) for “Recipient’s Age $\geq 10\text{a}$ ”. Descriptive analysis of the comparison groups for day 365(+/-20) after HSCT (class recipient’s age) is shown in Figure 32.

Day 365 (+/- 20) after TX

	Recipient's Age <10a	Recipients's Age $\geq 10\text{a}$
n	79	58
Max	1537	305
Min	1	0
Q1	153	19.5
Median	302	63
Q3	456	139.5
Mean	357.04	87.31
σ (std.dev)	314.40	81.71
variance	98849.18	6676.35
SEM	35.37	10.73
Kurtosis	6.54	3.02

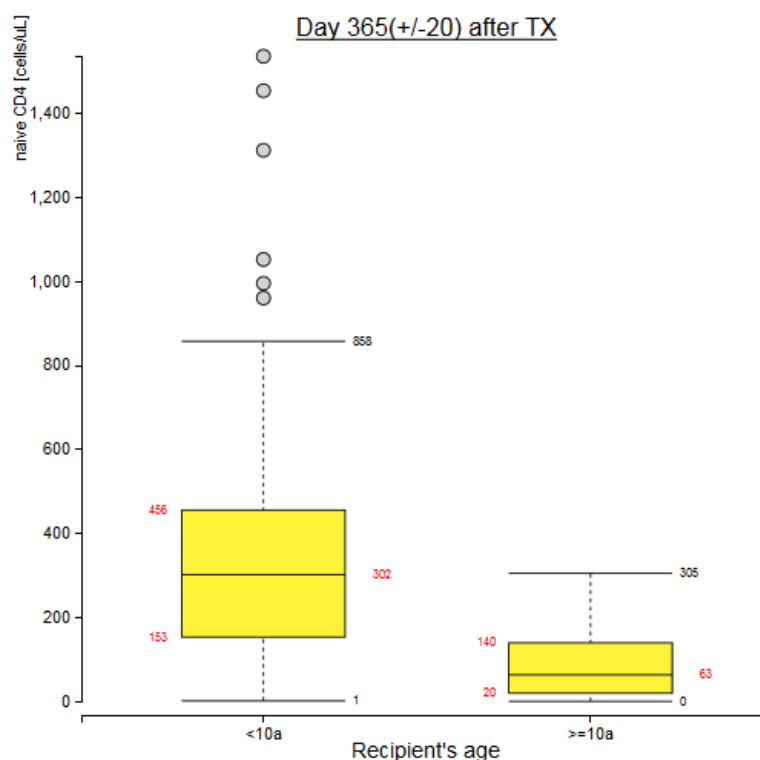


Figure 32: Recipient’s age comparison, 365 (+/-20) days after HSCT (TX)

3.3.2 Normal distribution analysis

In this section, the analysis of normal distribution is shown. The distribution of the data is crucial to choose the correct method for statistical testing. The figures shown in this section are exported from the program R. The headings of the subplots were afterwards added to the charts.

The histograms in Figure 33 (“graft source”), Figure 35 (“donor’s age”) and Figure 34 (“recipient’s age”) show the distribution of values (CD4+ T-cells/ μ l) at the testing days 30 (+/-5), 100(+/-10), 180(+/-20), 365(+/-20) after HSCT. All four histograms show a left-skewed data distribution. The minimum value is 0; no negative values are available. Therefore, no normal distribution is assumed and rank-scaled statistical test methods are applied.

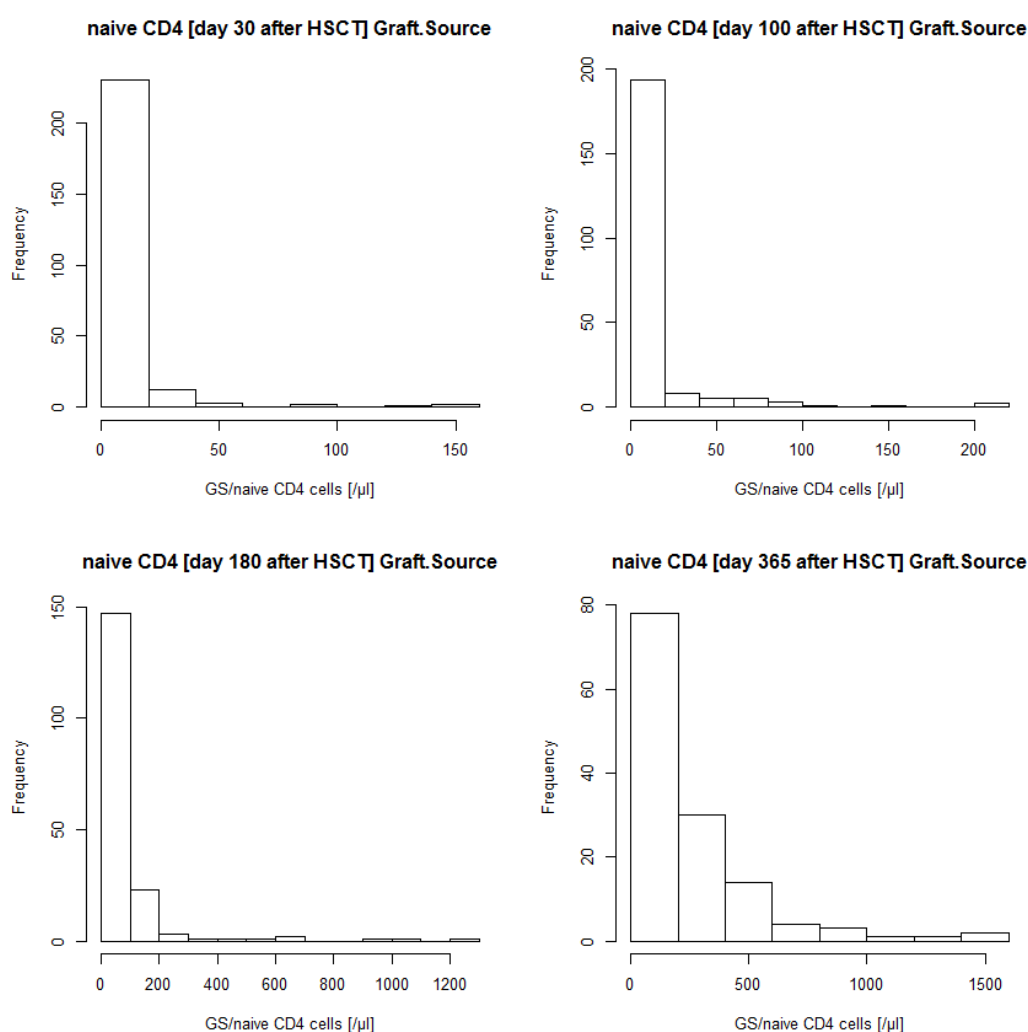


Figure 33: Distribution of data (category graft source comparison)

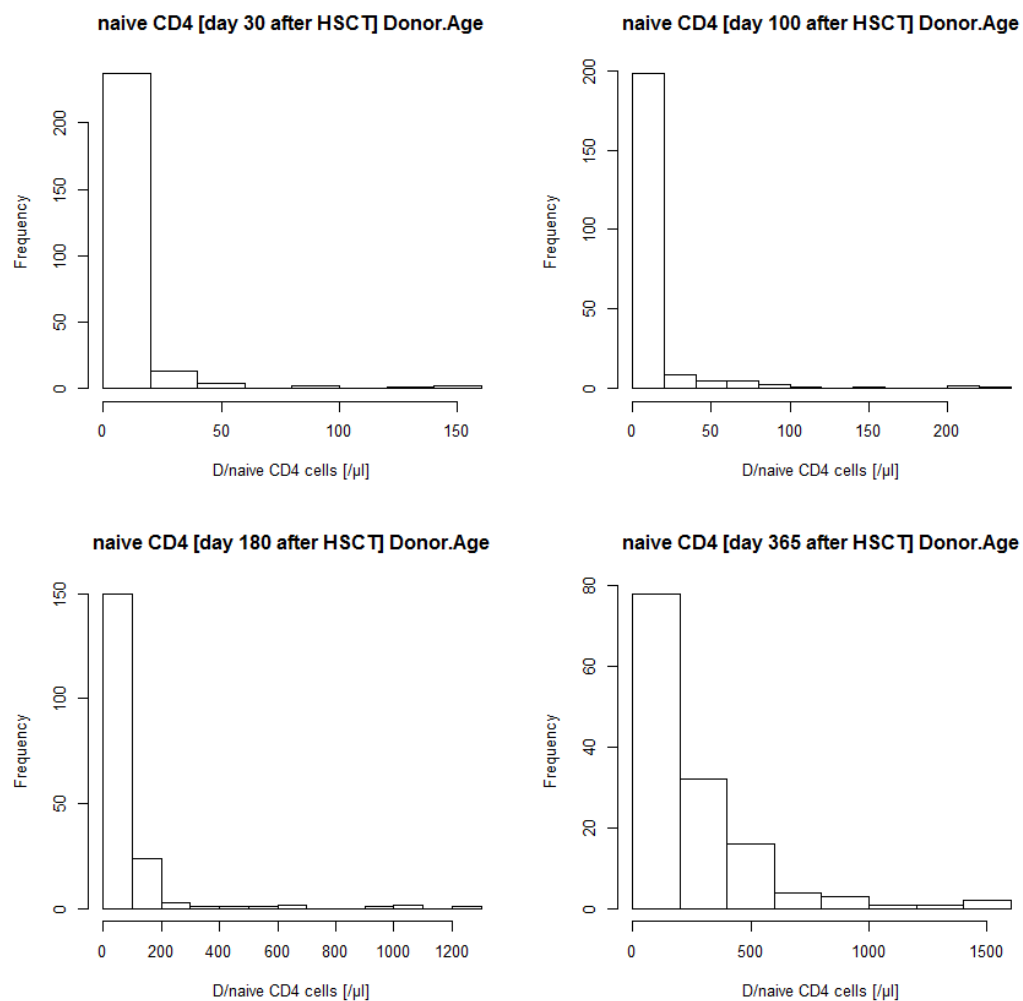


Figure 34: Distribution of data (category donor's age comparison)

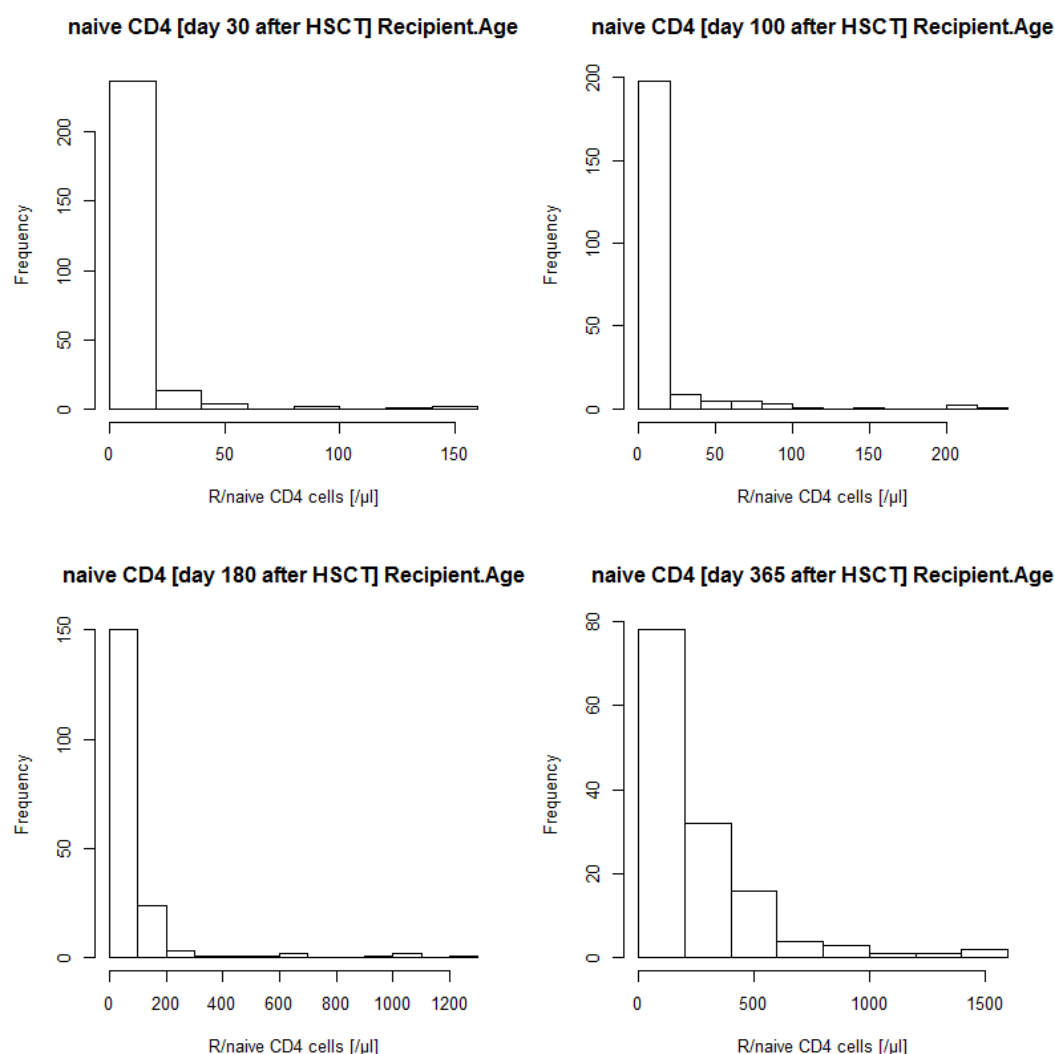


Figure 35: Distribution of data (category recipient's age comparison)

3.3.3 Comparison of groups

This section shows the result of statistical tests regarding comparison of groups. For three groups Kruskal-Wallis rank sum test and for two groups Wilcoxon rank sum test was applied. The displayed over-all probability value is adjusted by the Bonferroni correction factor of 4 (significant $p_{\text{corr}} < 0,05$). The related boxplots, including descriptive statistical analysis, can be found in section 3.3.1. p_{corr} is the Bonferroni-corrected p value. The related boxplots and descriptive statistic can be found in section 3.3.1.

Graft Source

The comparison of groups related to different graft sources (BM, PBSC, PBSC_CD3_depl) in relation to the cell count of naïve CD4+ T-cells, can be found in Table 10. A significant difference comparing the 3 groups could only be

detected on day 30 after HSCT (Kruskal-Wallis rank sum test, $p_{\text{corr}} = 0.029$), a post-hoc analysis showed a significant difference between BM₃₀ and PBSC_CD3_depl₃₀ (pairwise Wilcoxon rank sum test, $p(\text{BM vs. PBSC_CD3_depl}) = 0.011$). All other comparative tests showed no significant differences. The tests, regarding comparison of groups, with significant difference are highlighted in green.

*Table 10: Comparison of groups (graft source).
The related descriptive statistic and boxplots can be found in section 3.3.1.1*

days after HSCT	Test Method	BM	PBSC	PBSC_CD3_depl
30 (+/-5) Figure 21	Kruskal-Wallis rank sum test	over all $p_{\text{corr}} = 0.029$ $\chi^2 = 9.8785$		
	Bonferroni correction			
	pairwise Wilcoxon rank sum test	p(BM vs. PBSC) = 0.160 n.s.		
		p(BM vs. PBSC_CD3_depl) = 0.011		
100 (+/-10) Figure 22	Kruskal-Wallis rank sum test	over all $p_{\text{corr}} = 0.443$ n.s. $\chi^2 = 4.4023$		
	Bonferroni correction			
180 (+/-20) Figure 23	Kruskal-Wallis rank sum test	over all $p_{\text{corr}} = 0.156$ n.s. $\chi^2 = 6.4919$		
	Bonferroni correction			
365 (+/-20) Figure 24	Kruskal-Wallis rank sum test	over all $p_{\text{corr}} = 1$ n.s. $\chi^2 = 1.1175$		
	Bonferroni correction			

Donor's age

By analyzing the “Donor's age”, comparison between three groups result in a significant over-all difference on day 30 (Kruskal-Wallis rank sum test, $p_{\text{corr}} < 0.001$ sig.), 100 (Kruskal-Wallis rank sum test, $p_{\text{corr}} < 0.001$ sig.) and 180 (Kruskal-Wallis rank sum test, $p_{\text{corr}} = 0.001$ sig.) after HSCT. The post hoc analysis was significant different between groups on day 30 (pairwise Wilcoxon rank sum test; $p(<10a \text{ vs. } >20a) < 0.001$ sig., $p(10-20a \text{ vs. } >20a) < 0.001$ sig.), on day 100 (pairwise Wilcoxon rank sum test; $p(<10a \text{ vs. } >20a) < 0.001$ sig., $p(10-20a \text{ vs. } >20a) < 0.001$ sig.) and on day 180 (pairwise Wilcoxon rank sum test; $p(<10a \text{ vs. } >20a) < 0.001$ sig, $p(10-20a \text{ vs. } >20a) < 0.033$ sig.). For details see Table 11. The comparison of groups resulted in significant differences as highlighted in green.

Table 11 Comparison of groups (Donor's age)
The related descriptive statistic and boxplots can be found in section 3.3.1.2

day after HSCT	Test Method	Donor's age <10a	Donor's age 10a – 20a	Donor's age > 20a
30 (+/-5) Figure 25	Kruskal-Wallis rank sum test	over all $p_{\text{corr}} < 0.001$		
	Bonferroni correction	$\chi^2 = 46.684$		
	pairwise Wilcoxon rank sum test	$p(<10a \text{ vs. } 10-20a) = 0.85 \text{ n.s.}$		
		$p(<10a \text{ vs. } >20a) < 0.001$		
100 (+/-10) Figure 26	Kruskal-Wallis rank sum test	over all $p_{\text{corr}} < 0.001$		
	Bonferroni correction	$\chi^2 = 43.843$		
	pairwise Wilcoxon rank sum test	$p(<10a \text{ vs. } 10-20a) = 0.31 \text{ n.s.}$		
		$p(<10a \text{ vs. } >20a) < 0.001$		
180 (+/-20) Figure 27	Kruskal-Wallis rank sum test	over all $p_{\text{corr}} = 0.001$		
	Bonferroni correction	$\chi^2 = 16.537$		
	pairwise Wilcoxon rank sum test	$p(<10a \text{ vs. } 10-20a) = 0.093 \text{ n.s.}$		
		$p(<10a \text{ vs. } >20a) < 0.001$		
365 (+/-20) Figure 28	Kruskal-Wallis rank sum test	over all $p_{\text{corr}} = 1 \text{ n.s.}$		
	Bonferroni correction	$\chi^2 = 1.5003$		

Recipient's age

In the category “recipient's age”, a significant difference between groups “Recipient's age <10a” and “Recipient's age $\geq 10a$ ” was seen on days 180 (Wilcoxon rank sum test, $p_{\text{corr}} < 0.001$ sig.) and 365 (Wilcoxon rank sum test, $p_{\text{corr}} < 0.001$ sig.) after HSCT. For details see Table 12. The tests, regarding comparison of groups, with significant difference, are highlighted in green.

Table 12: Comparison of groups (Recipient's age)
The related descriptive statistic and boxplots can be found in section 3.3.1.3

days after HSCT	Test Method	Recipient's age <10a	Recipient's age $\geq 10a$
30 (+/-5) Figure 29	Wilcoxon rank sum test	over all $p_{\text{corr}} = 1 \text{ n.s.}$	
	Bonferroni correction		
100 (+/-10) Figure 30	Wilcoxon rank sum test	over all $p_{\text{corr}} = 0.166 \text{ n.s.}$	
	Bonferroni correction		
180 (+/-20) Figure 31	Wilcoxon rank sum test	over all $p_{\text{corr}} < 0.001$	
	Bonferroni correction		
365 (+/-20) Figure 32	Wilcoxon rank sum test	over all $p_{\text{corr}} < 0.001$	
	Bonferroni correction		

3.3.4 Correlations

This section shows the result of a Spearman's rank correlation test of naïve CD4+ T-cells counts/ μl between the groups recipient's age and donor's age. At the last part of this chapter a scatterplot matrix (SPLOM), representing all

possible pairwise combinations of all three attributes (donor's age, recipient's age, count of CD4+ T-cells), is shown. All graphs are designed for measurements on days 30 (+/-5), 100 (+/-10), 180 (+/-20) and 365 (+/-20) after HSCT.

The calculated probability value had to be adjusted by the factor 4 (Bonferroni correction) due to four tests on the same dataset. The corrected value as shown as $p(\text{corr.bfr})$.

3.3.4.1 Recipient's age

A correlation analysis association between naïve CD4+ T-cells count and recipient's age was performed. The results are shown in Figure 36. The correlation between naïve CD4 T-cell count and recipient's age was shown to be significant only on days 180 and 365 after HSCT ($p(\text{corr.bfr})_{180} < 0.001$, $p(\text{corr.bfr})_{365} < 0.001$). In contrast, the correlation coefficient is negative on days 100 (Spearman $\rho_{100} = -0.135$, $p(\text{corr.bfr})_{100} = 0.1752$ n.s.), 180 (Spearman $\rho_{180} = -0.438$, $p(\text{corr.bfr})_{180} < 0.001$ sig.) and 365 (Spearman $\rho_{365} = -0.645$, $p(\text{corr.bfr})_{100} < 0.001$ sig.) after HSCT.

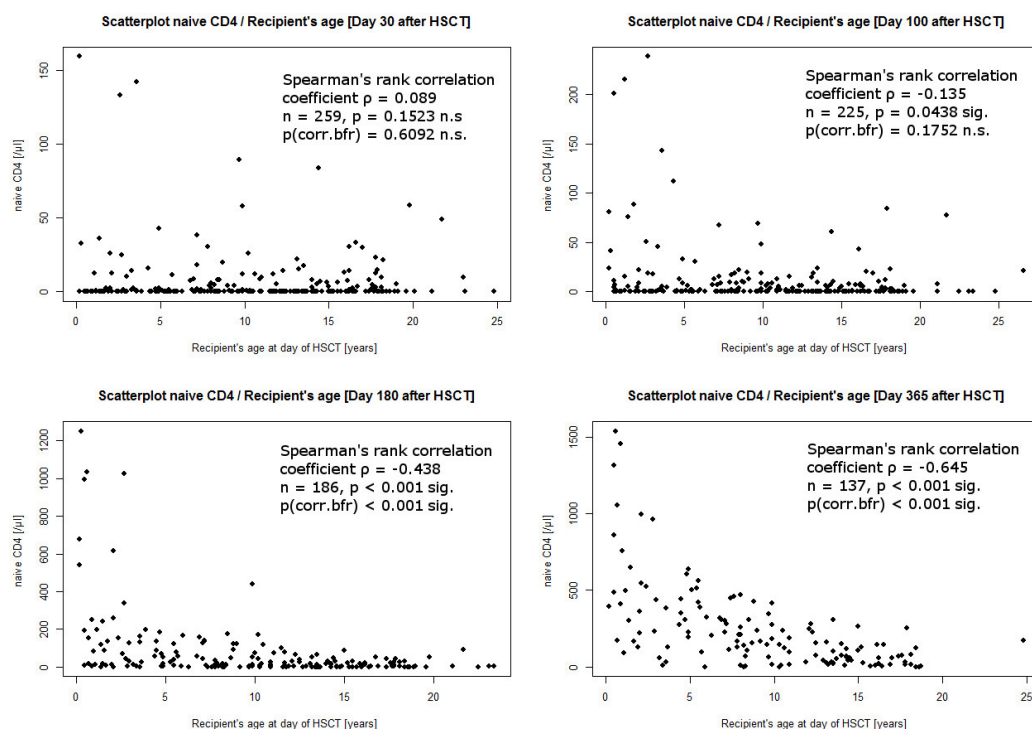


Figure 36: Spearman's rank correlation (category recipient's age),

3.3.4.2 Donor's age

A correlation analysis between naïve CD4+ T-cell count and donor's age was performed. The results are shown in Figure 36. The correlation between naïve CD4 cell count and donor's age was only significant on days 30, 100 and 180

after HSCT ($p(\text{corr.bfr})_{30} < 0.001$, $p(\text{corr.bfr})_{100} < 0.001$, $p(\text{corr.bfr})_{365} < 0.001$), respectively. In addition, there is an inverse correlation between the count of naïve CD4+ T-cells and recipient's age at all time points of analyses (Spearman $\rho_{30} = -0.422$, $p(\text{corr.bfr})_{30} < 0.001$ sig., $\rho_{100} = -0.432$, $p(\text{corr.bfr})_{100} < 0.001$ sig., $\rho_{180} = -0.340$, $p(\text{corr.bfr})_{180} < 0.001$ sig., $\rho_{365} = -0.077$, $p(\text{corr.bfr})_{100} = 1$ n.s.) after HSCT.

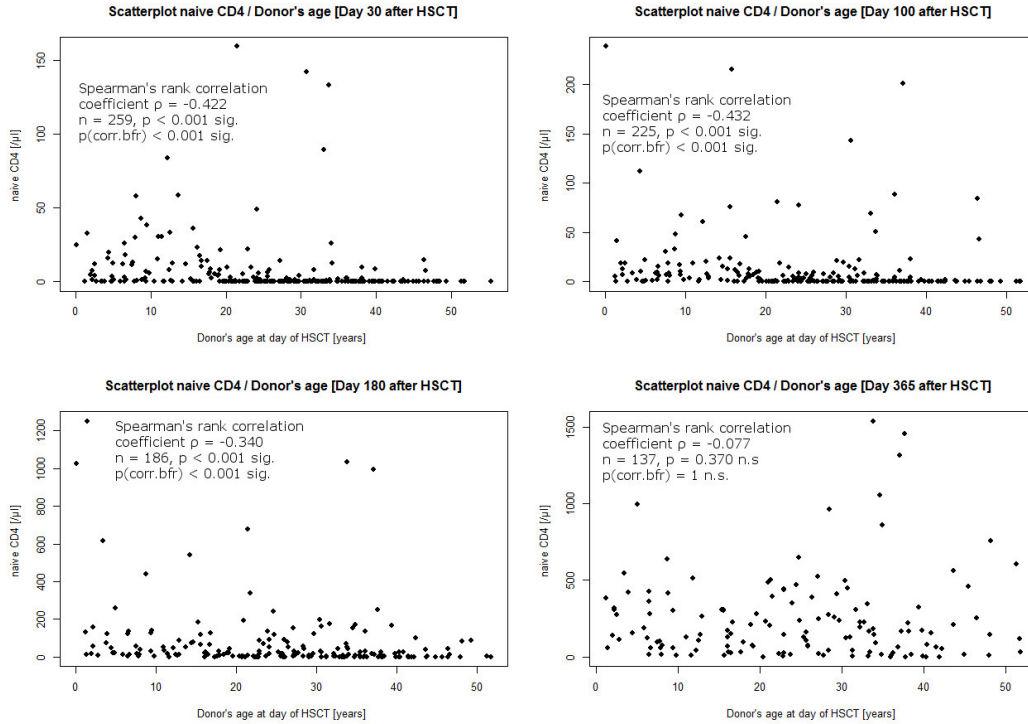


Figure 37: Spearman's rank correlation (category donor's age)

Comparing the donor's age and with recipient's age, no significant correlation was seen (Spearman $\rho_{30} = 0.020$, $p(\text{corr.bfr})_{30} = 1$ n.s., $\rho_{100} = 0.047$, $p(\text{corr.bfr})_{100} = 1$ n.s., $\rho_{180} = 0.025$, $p(\text{corr.bfr})_{180} = 1$ n.s., $\rho_{365} = 0.009$, $p(\text{corr.bfr})_{100} = 1$ n.s.).

The set of scatterplot matrices including donor's age and recipient's age is shown in Figure 38, Figure 39, Figure 40 and Figure 41. The sub-plots recipient_age [year] vs. donor_age [years] illustrate the scattering and the small correlation coefficient.

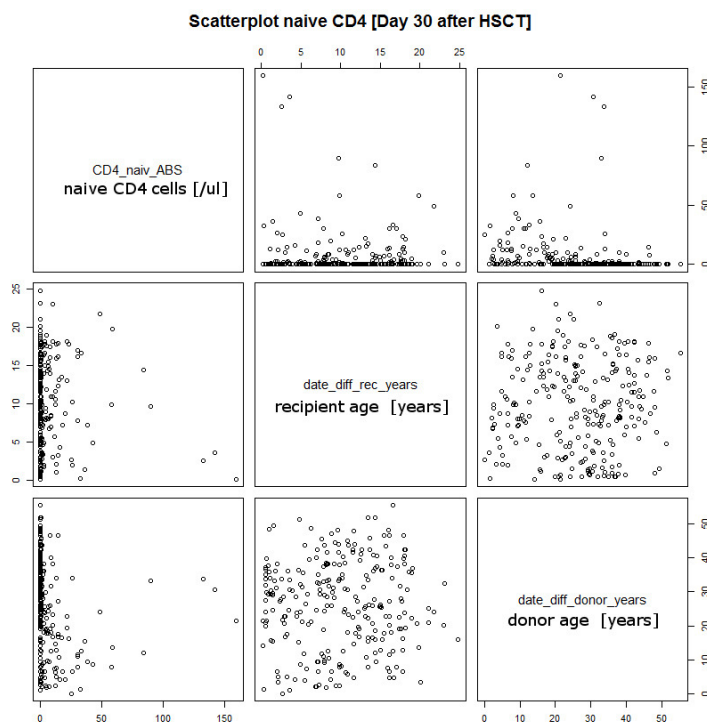


Figure 38: SPLOM recipient's age, donor's age, naïve CD4 count, day 30 (+/-5) after HSCT

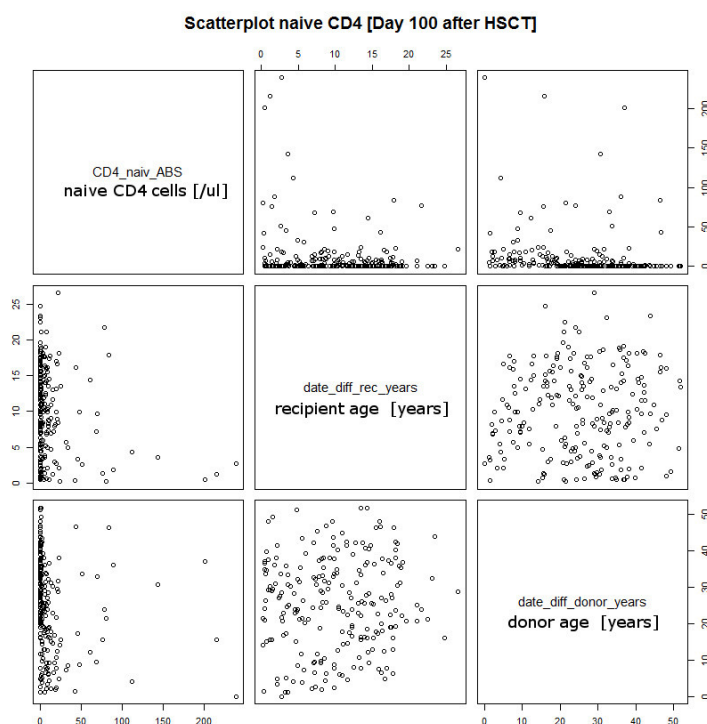


Figure 39: SPLOM recipient's age, donor's age, naïve CD4 count, day 100 (+/-10) after HSCT

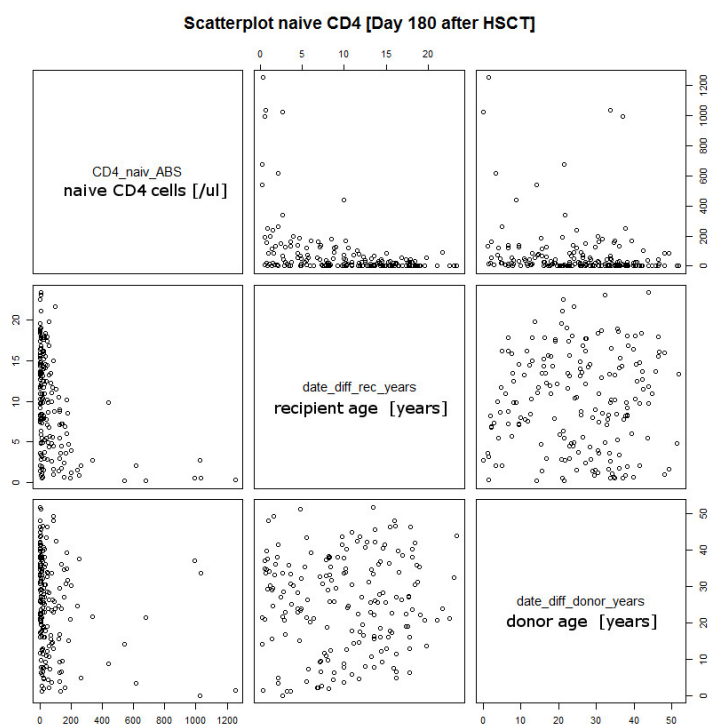


Figure 40: SPLOM recipient's age, donor's age, naïve CD4 count, day 180 (+/-20) after HSCT

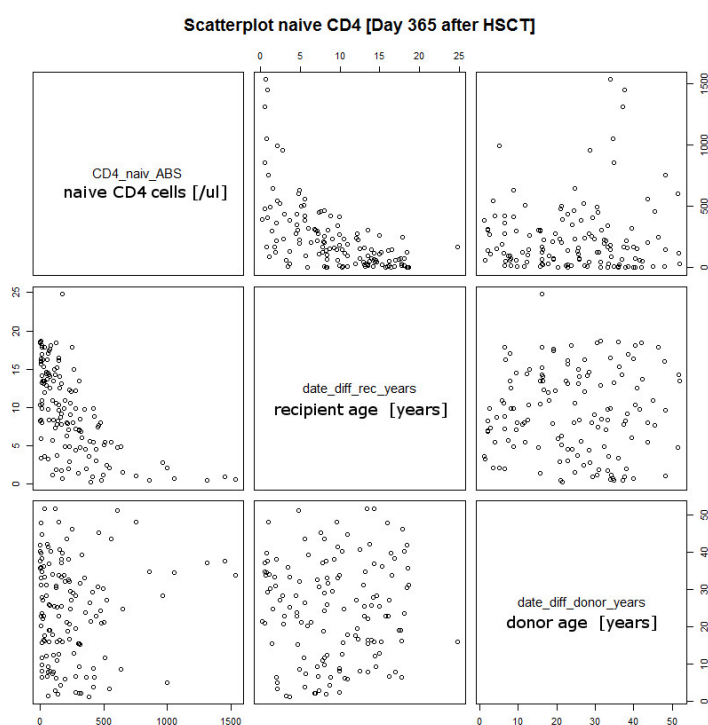


Figure 41: SPLOM recipient's age, donor's age, naïve CD4 count, day 365(+/-20) after HSCT

4 Discussion

The first aim of this thesis was the conversion of collected and stored historical flow cytometry data into a rational database including web-based visualization to depict cell counts via an interactive line chart. Secondly, an explorative statistical testing of the collected data, focusing on engraftment characteristics was performed. The involvement of the adaptive immune system (naïve CD4⁺ T-cells) during HSCT was analyzed. Thereby, the focus was on the influence of the graft source (bone marrow, peripheral blood stem cells and CD3-depleted peripheral blood stem cells), donor's age, and the recipient's age.

The following sections discuss the results of the work packages give an answer to the research questions (defined in section 1.3.1).

4.1 Data processing

WP1 Feasibility: Transfer of existing records of flow cytometry analyses data into a platform, which allows an easier data access and advanced analyzability. **Fulfilled successfully**

The data scraping process of historical patient's flow cytometry data could be successfully fulfilled. An essential part was the manual preparation (section 2.4.3) and check of the human-readable original files. The software (`parse_xls`, see section 2.4.5.2) was designed for the transition and data scraping procedure but could not detect all possible abnormalities in the original data structure. Therefore, results of cell-subset and remarks had to be assigned to the correct columns. Furthermore, missing information about the type of specimen and analysis method had to be redefined. This manual checking was successfully performed on 552 files. Special cases, such as two transplantations in a single file or additional values, were marked accordingly. These files were corrected after consultation with the head of the department. The software `parse_xls` reports to the user any change of data before transmitting to the database. Therefore, overlooked abnormalities in the original Excel files and misinterpretations were reported to user by the program.

The implemented MariaDB database builds an open source stable system, with access control, within the CCRI network. It has the benefit of a daily backup service and provides the possibility of different queries over all 26492 flow cytometry analysis datasets and the related properties of transplantation, graft

source and the underlying disease. In contrast, a manual scientific evaluation of more than 400 separate Excel files would be very time consuming. Therefore, this novel database provides a source for better and faster analyzability and facilitates extraction of relevant data for answering scientific questions related to retrospective analysis of engraftment of cell-subsets after pediatric HSCT.

4.1.1 Limitation and outlook

Access was limited to the clinical dataset of the documentary file `STAMM1C.xls`. Therefore only a reduced amount of data, potentially influencing the engraftment process, could be captured and stored in the database.

A further limitation of this data processing was the fact that values were manually copied multiple times from the analysis file of the flow cytometer into the Excel tables. Also, data was defined as below ($< [\text{value}]$) or over ($> [\text{value}]$) a certain value (e.g. <0.1) for the database entry. This skewing, as well as rounding, must be seen as a bias.

As the database bears opportunities for data exploration and analysis, the database should be used in daily routine. Therefore an additional interface for transmission of the data between the sources of generation (e.g. flow cytometer, PCR chimerism analysis laboratory, FISH chimerism laboratory) must be implemented for central storage and data integrity. The training of the working staff for using this state-of-the-art technology and a usability testing remains to be done.

A further benefit of this structured database is the possibility to generate reports for clinical requesters.

4.2 Interactive web-based visualization

WP2 Feasibility: Development of a web-based interactive visualization tool for flow cytometry data with the option to filter results according to graft source, human leukocyte antigen (HLA) mismatch and number of transplanted CD34+ and CD3+ cells. **WP fulfilled successfully.**

Interactive visualization of medical data supports analysts and support the physicians to get a fast overview on complex data of patients [44]. Therefore, in this thesis, flow cytometry data was visualized by a web-based line-chart with interactive features. When exploring time-oriented data for discovery of patterns and trends, it is important to implement two main features. First, filtering in the attribute data and second, panning and zooming in the attribute time [45]. Both features were technically implemented (see Figure 14) and can provide an easy-accessible graphical representation of different cell lines and their engraftment

characteristics. The categories “Human Leucocyte Antigen (HLA) mismatch between donor and recipient”, “donor’s sex”, “origin of graft”, and “count of CD34+ cells and CD3+ T-cells within the graft” can be filtered on the website. The effect of the applied filter is a displayed list of patients, which fulfills the selection requests. The visualization opens up new opportunities to generate new hypothesis for engraftment characteristics and correlations between properties of patient, type of transplantation and patterns of engraftment. Additional selection and filtering features of certain lines, representing different cells, supports the analyst in order to gain a more detailed view of cell counts over time. Because of the web-based implementation, the data can be accessed within the network of the St. Anna Children’s Cancer Research Institute. Therefore, the installation of additional software is not required. To ensure restricted access to patient’s data, an obligatory previous registration process was implemented.

4.2.1 Limitation and outlook

The scope of the thesis was to visualize patient’s flow cytometry data. In a future approach an overlay of patient’s line charts with similar characteristics could be interesting to group analysis over time in one time chart. This would facilitate finding patterns and differences between groups and generate new hypothesis to be proofed.

The current development only allows the representation of web-based boxplots and their descriptive statistics in the browsers Firefox and Chrome. For better accessibility, additional browsers should be tested and the functionality evaluated.

Limitation of this part of the development is the missing encryption of data exchange between server and client which has to be implemented, if this software should be accessible outside the CCRI’s network.

4.3 Statistical analyses

RQ1: Do graft source, age of donor and age of recipient influence the amount of naïve CD4 T cells on days 30, 100, 180, 365 after allo-HSCT?

The transition of historical data and the possibility of continuous central storage open doors for different approaches of evaluations and analyses. To demonstrate an area of application, a retrospective explorative data analysis of naïve CD4+ T-cells was performed. These analyses build just a first step to answer clinical relevant questions and must be interpreted carefully, as only a limited set of statistical methods was applied. Further investigation on data and its

interdependences between patients, their diseases and the transplantation setting, as well as, on statistical tests methods is absolutely necessary.

RQ1.1 Statistical Analysis: Does the graft source (BM, PBSC, CD3 depleted PBSC) influence the engraftment process (absolute cell count) of naïve CD4+ T-cells after allo-HSCT? yes, with limitations, partially

Based on clinical observation at the transplantation unit at the St. Anna Children's a possible influence of different graft sources on naïve CD4+ T-cells counts was assumed. It is known, that fast recovery of naïve CD4+ cells may contribute to a reduced mortality after HSCT [24, 25]. Therefore, naïve CD4+ T-cell counts at different time points after HSCT (day 30(+/-5), 100(+/-10), 180(+/-20), 356(+/-20)) of patients transplanted with either BM, PBSCs or CD3+ T-cells depleted PBSC, were compared with each other. As expected, the size of the groups was not equal and distributions of additional characteristics (e.g. disease, age, methods for conditioning, etc.) between groups (BM, PBSC, CD3 depleted PBSC) were not examined. The values in groups were not normally distributed (see Figure 33).

A significant difference between all groups could be only observed on day 30 after HSCT (Kruskal-Wallis rank sum test with Bonferroni correction, over all $p_{\text{corr}} = 0.029$ sig.), a pairwise sub-group comparison resulted in a significant difference between "BM" and "PBSC_CD3_depl" (pairwise Wilcoxon rank sum test, $p(\text{BM}, \text{PBSC_CD3_depl}) = 0.011$ sig.) (Table 10). A trend was observed, that the mean cell count of naïve CD4+ T-cells was higher in the group "BM" in comparison to "PBSC" and "PBSC_CD3_dep" (see Figure 21, Figure 22 and Figure 23), with the exception of day 365 (Figure 24). Concerning the limitations, the group size of "PBSC_CD3_dep" was small ($n_{\text{PBSC_CD3_delp_365}} = 7$), compared to other groups ($n_{\text{PBSC_CD3_delp_30}} = 14$, $n_{\text{PBSC_CD3_delp_100}} = 17$, $n_{\text{PBSC_CD3_delp_180}} = 18$).

RQ1.2 Statistical Analysis: Does the engraftment process (absolute count) of naïve CD4+ T-cells correlate with the age of the donor? yes, with limitations, partially

Based on the assumption that donor's age influence the count of naïve CD4+ T-cells [24], three different groups of donors with the age <10a, 10a-20a and >20a, were analyzed. The number of patients in these groups was unequal, because usually most of the stem cell donors are older than 20 years. A distribution analysis of additional characteristics (e.g. disease, methods for conditioning, etc.) between groups (<10a, 10a-20a and >20a) was not performed. The values for counts of naïve CD4+ did not follow a normal distribution (see Figure 34).

Croup comparison, based on Kruskal-Wallis rank sum test with Bonferroni correction, showed a significant over-all difference between the three groups (<10a, 10a-20a and >20a) on days 30 (over all $p_{\text{corr}} < 0.001$ sig), 100 (over all

$p_{\text{corr}} < 0.001$ sig), 180 (over all $p_{\text{corr}} = 0.001$ sig) after HSCT (see Table 11). Within these three groups of donor's age, a pairwise group comparison, based on Wilcoxon rank sum test, showed a significant difference between groups (<10a, >20a) and (10-20a, >20a). Concerning the mean values, the group with donors older than 20 years has the lowest median CD4+ T-cell count at measurement points of day 30, 100, 180 after HSCT.

Concerning the Spearman's rank correlation coefficient, a significant inverse correlation ($p(\text{corr.bfr}) < 0.001$ sig.) between the absolute count of CD4+ T-cells and the age of donor on days 30 ($p = -0,422$, moderate), 100 ($p = -0,432$, moderate) and 180 ($p = -0,340$, weak) after HSCT were seen (Figure 37). The correlation on day 360 is very weak ($p = -0,077$), and not significant. This supports the hypothesis, that young donors result in a higher number of naïve CD4+ T-cell counts on days 30, 100, 180 in patients after HSCT. Furthermore the scatter plot matrices (Figure 38, Figure 39, Figure 40 and Figure 41) suggest, graphically, no correlation between donor's age and recipient's age.

The correlation between donor's age and recipient's age, as well as multivariable tests, remains to be examined.

RQ1.3. Statistical Analysis: Does the engraftment process (absolute count) of naïve CD4+ T-cells correlate with the age of the recipient?
yes, with limitations, partially

Concerning the influence of recipient's age on cell counts of naïve CD4+ T-cells, two groups of recipients (<10 years) and (≥ 10 years) of age, was performed, without analysis of distribution of sub-characteristics (e.g. disease, methods for conditioning etc.) of both groups. The size of group was approximately equal, but without assurance of normal distribution of naïve CD4 T-cell values (see Figure 35: Distribution of data (category recipient's age comparison). The Wilcoxon rank sum test with Bonferroni correction, resulted in in a significant difference at measurement points of day 180 and day 365 (see Table 12). The median absolute cell count of naïve CD4+ T-cells of the test group "recipient's age <10a" was higher than in group "recipient's age $\geq 10a$ " on days 100, 180 and 365 after HSCT. The Spearman's rank correlation resulted in a significant negative correlation coefficient on day 180 ($p = -0,438$, moderate, $p(\text{corr.bfr}) < 0.001$ sig.) and 365 ($p = -0,645$, strong, $p(\text{corr.bfr}) < 0.001$ sig.).

These observation suggest, that younger (<10a) patients, transplanted at the St. Anna Children's hospital, trend to have a higher count of naïve CD4+ T-cells after day 180 after HSCT, than older ones ($\geq 10a$).

The correlation between recipient's age and donor's age, as well as multivariable tests remains to be examined.

WP3 Feasibility: Development of a web-based tool for descriptive statistics and distribution of data for questions RQ1.1-RQ1.3. fulfilled, with limitations

It has been shown, that with the support of the java class D3.js, SVG based boxplots can be designed, to get a fast overview about data distributions in different groups. This facilitates the analysts to find patterns easier [21] and create working hypothesis, which have to be examined. An additional table, next to the boxplots, demonstrates the descriptive statistic. The displayed values like standard deviation and variance have to be interpreted carefully, as distribution of values is not normal.

A usability test about technology acceptance, remains to be done to evaluate benefits of this application and its data representation [46]. The application does not fully work when accessed with the Internet Explorer. This issue needs to be addressed.

A further development of interactivity could be the implementation of exploding boxplots [47]. This visualization technique provides a good overview about the data distribution and the quartiles, but also a more detailed view, if necessary (see Figure 42).

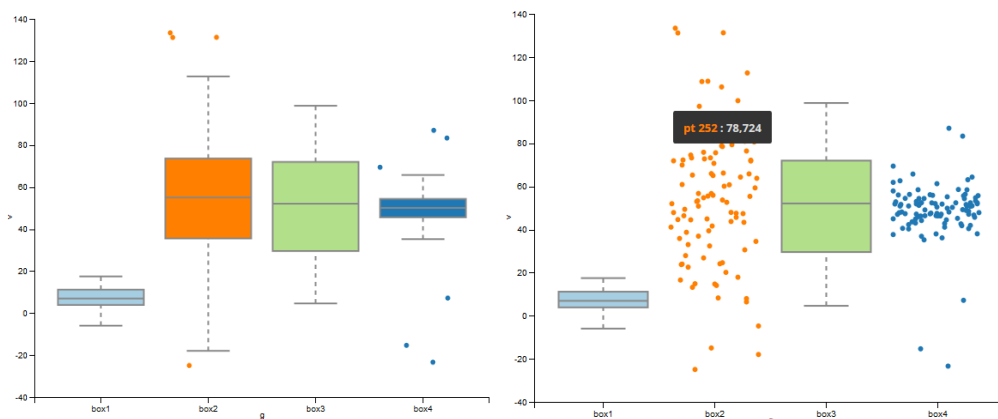


Figure 42: Exploding boxplots (adopted from [47]). Interaction facilitates a detailed distribution on data points. A click on the boxplot shows the data points.

4.3.1 Limitations

General limitation of the statistical analyses is that it is based on a retrospective analysis, where influencing factors are already set and cannot be designed prospectively in an experiment. Furthermore, the applied correction of Bonferroni, due to the fact, that multiple tests were performed on the same dataset, is very conservative, potentially including wrong negative significance levels [48].

The basic explorative statistical analysis in this paper shows an example for data analysis, based on the novel database. To achieve clinical relevance, additional factors for statistical analysis must be considered!

5 Conclusion

The aim of this thesis was to design a framework to increase analyzability of flow cytometry data. Therefore, historical reports were analyzed and data scraping was performed. A web-based visualization technique shall support analysts to find patterns, while performing interactive data sniffing in free adjustable time-oriented line charts.

The explorative statistical tests and correlations have shown the possibilities of a structured database system. Therefore this preparation of data can be concluded as useful for further investigations of flow cytometry data related to cell engraftment after pediatric HSCT.

Trends have shown, that donor's and recipient age might influence the cell count of naïve CD4+ T-cells during engraftment after HSCT. However, clinical relevance and correlations between different cells of the adaptive immune system and the occurrence of e.g. infections remain to be investigated.

At the St. Anna's Children Cancer Research institute, for researches and clinical related projects the integration of the database and the according visualization and statistical tools are crucial to drive "Digital Healthcare" forward.

Literature

- [1] B. Lindner/ÖGHO. "Transplantdaten der letzten 10 Jahre" [email]. Sender: ASCTR@i-med.ac.at [Received: 06/07/2017]
- [2] H. Pichler, V. Witt, E. Winter, H. Boztug, E. Glogova, U. Potschger, *et al.*, "No impact of total or myeloid Cd34+ cell numbers on neutrophil engraftment and transplantation-related mortality after allogeneic pediatric bone marrow transplantation," *Biol Blood Marrow Transplant*, vol. 20, pp. 676-83, May 2014.
- [3] D. Gotz, F. Wang, and A. Perer, "A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data," *J Biomed Inform*, vol. 48, pp. 148-59, Apr 2014.
- [4] S. Silbernagl and A. Despopoulos, *Taschenatlas Physiologie*. Stuttgart: Thieme, 2012.
- [5] F. Lang and P. Lang, *Basiswissen Physiologie*. Berlin, Heidelberg: Springer Medizin Verlag Heidelberg, 2007.
- [6] M. P. Gawaz, *Blood platelets : physiology, pathophysiology, membrane receptors, antiplatelet drugs, coronary heart disease, stroke, peripheral arterial disease : 47 tables*. Stuttgart [u.a.: Thieme, 2001.
- [7] J. Zierk, F. Arzideh, T. Rechenauer, R. Haeckel, W. Rascher, M. Metzler, *et al.*, "Age- and sex-specific dynamics in 22 hematologic and biochemical analytes from birth to adolescence," *Clin Chem*, vol. 61, pp. 964-73, Jul 2015.
- [8] A. Görgens, S. Radtke, M. Möllmann, M. Cross, J. Dürig, Peter A. Horn, *et al.*, "Revision of the Human Hematopoietic Tree: Granulocyte Subtypes Derive from Distinct Hematopoietic Lineages," *Cell Reports*, vol. 3, pp. 1539-1552.
- [9] N.-p. Weng, "Aging of the Immune System: How Much Can the Adaptive Immune System Adapt?," *Immunity*, vol. 24, pp. 495-499, 5// 2006.
- [10] M. Häggström. "Hematopoiesis (human) diagram" [Online]. Available: https://commons.wikimedia.org/wiki/File:Hematopoiesis_simple.svg [Accessed: 16-05-2017]

- [11] J. Dmytrus, S. Matthes-Martin, H. Pichler, N. Worel, R. Geyeregger, N. Frank, *et al.*, "Multi-color immune-phenotyping of CD34 subsets reveals unexpected differences between various stem cell sources," *Bone Marrow Transplant*, vol. 51, pp. 1093-100, Aug 2016.
- [12] R. Hoffman, *Hematology : basic principles and practice*, 6th ed. Philadelphia, PA: Saunders/Elsevier, 2013.
- [13] L. Gattinoni, D. E. Speiser, M. Lichterfeld, and C. Bonini, "T memory stem cells in health and disease," *Nat Med*, vol. 23, pp. 18-27, Jan 06 2017.
- [14] M. Berard and D. F. Tough, "Qualitative differences between naïve and memory T cells," *Immunology*, vol. 106, pp. 127-138, 04/17 2002.
- [15] E. Nowak, "STAMM1C für Ewa und Susi.xls," ed. St. Anna Kinderspital, 2017.
- [16] T. A. Ghaleb, M. A. Mohammed, and E. Ramadan, "Automated analysis of flow cytometry data: a systematic review of recent methods," in *2016 2nd International Conference on Open Source Software Computing (OSSCOM)*, 2016, pp. 1-7.
- [17] G. Fritsch, P. Buchinger, and D. Printz, "Use of flow cytometric CD34 analysis to quantify hematopoietic progenitor cells," *Leuk Lymphoma*, vol. 10, pp. 443-51, Aug 1993.
- [18] J. Pichler, D. Printz, D. Scharner, D. Trbojevic, J. Siekmann, and G. Fritsch, "Improved flow cytometric method to enumerate residual cells: minimal linear detection limits for platelets, erythrocytes, and leukocytes," *Cytometry*, vol. 50, pp. 231-7, Aug 15 2002.
- [19] F. Khan, A. Agarwal, and S. Agrawal, "Significance of chimerism in hematopoietic stem cell transplantation: new variations on an old theme," *Bone Marrow Transplant*, vol. 34, pp. 1-12, 05/24/online 2004.
- [20] J. L. Diez-Martin, P. Llamas, J. Gosalvez, C. Lopez-Fernandez, N. Polo, M. S. de la Fuente, *et al.*, "Conventional cytogenetics and FISH evaluation of chimerism after sex-mismatched bone marrow transplantation (BMT) and donor leukocyte infusion (DLI)," *Haematologica*, vol. 83, pp. 408-15, May 1998.
- [21] T. Munzner, *Visualization analysis and design*. Boca Raton: CRC Press, 2014.

- [22] M. Tanaka. "C3.js | D3-based reusable chart library" [online]. Available: <http://c3js.org/> [Accessed: June 12, 2017]
- [23] A. Rind, T. D. Wang, W. Aigner, S. Miksch, K. Wongsuphasawat, C. Plaisant, *et al.*, "Interactive Information Visualization to Explore and Query Electronic Health Records," *Foundations and Trends® in Human-Computer Interaction*, vol. 5, pp. 207-298, 2013.
- [24] E. Azuma, M. Hirayama, H. Yamamoto, and Y. Komada, "The role of donor age in naive T-cell recovery following allogeneic hematopoietic stem cell transplantation: the younger the better," *Leuk Lymphoma*, vol. 43, pp. 735-9, Apr 2002.
- [25] J. D. Goldberg, J. Zheng, R. Ratan, T. N. Small, K. C. Lai, F. Boulad, *et al.*, "Early recovery of T-cell function predicts improved survival after T-cell depleted allogeneic transplant," *Leuk Lymphoma*, vol. 58, pp. 1859-1871, Aug 2017.
- [26] S. Maury, J. Y. Mary, C. Rabian, M. Schwarzingier, A. Toubert, C. Scieux, *et al.*, "Prolonged immune deficiency following allogeneic stem cell transplantation: risk factors and complications in adult patients," *Br J Haematol*, vol. 115, pp. 630-41, Dec 2001.
- [27] B. N. Savani, S. Mielke, K. Rezvani, A. Montero, A. S. Yong, L. Wish, *et al.*, "Absolute lymphocyte count on day 30 is a surrogate for robust hematopoietic recovery and strongly predicts outcome after T cell-depleted allogeneic stem cell transplantation," *Biol Blood Marrow Transplant*, vol. 13, pp. 1216-23, Oct 2007.
- [28] S. Servais, E. Lengline, R. Porcher, M. Carmagnat, R. Peffault de Latour, M. Robin, *et al.*, "Long-term immune reconstitution and infection burden after mismatched hematopoietic stem cell transplantation," *Biol Blood Marrow Transplant*, vol. 20, pp. 507-17, Apr 2014.
- [29] Ethics Committee of the Medical University of Vienna. "Register for approved studies" [online]. Available: <https://ekmeduniwien.at/core/catalog/2017/> [Accessed: 23/05/2017]
- [30] wikiHow. "Ein sicheres Login Skript mit PHP und MySQL erstellen" [online]. Available: <http://de.wikihow.com/Ein-sicheres-Login-Skript-mit-PHP-und-MySQL-erstellen> [Accessed: 06/02/2017]
- [31] Microsoft. "Microsoft OLE DB" [online]. Available: [https://msdn.microsoft.com/en-us/library/ms722784\(v=vs.85\).aspx](https://msdn.microsoft.com/en-us/library/ms722784(v=vs.85).aspx) [Accessed: 06 June, 2017]

- [32] J. F.-. JQuery.Org. "jQuery" [online]. Available: <https://jquery.com/> [Accessed: June 12 2017]
- [33] C. Brewer. "ColorBrewer: Color Advice for Maps" [online]. Available: <http://colorbrewer2.org/> [Accessed: June 13, 2017]
- [34] J. Grubert. "D3.js Boxplot with Axes and Labels" [online]. Available: <http://bl.ocks.org/jensgrubert/7789216> [Accessed: June 13, 2017]
- [35] J. W. Tukey, *Exploratory data analysis*. Reading, Mass.: Addison-Wesley, 1993.
- [36] M. Hollander and D. A. Wolfe, "Kruskal-Wallis Rank Sum Test," in *Nonparametric statistical methods*, ed New York: John Wiley & Sons, 1973, pp. 115–120.
- [37] ETH_Zürich. "R: Kruskal-Wallis Rank Sum Test". Available: <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/kruskal.test.html> [Accessed: June 20, 2017]
- [38] M. Hollander and D. A. Wolfe, "Wilcoxon Rank Sum and Signed Rank Tests," in *Nonparametric statistical methods*, ed New York: John Wiley & Sons, 1973, pp. 27–33.
- [39] ETH_Zürich. "R: Wilcoxon Rank Sum and Signed Rank Tests". Available: <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/wilcox.test.html> [Accessed: June 20, 2017]
- [40] J. H. McDonald, *Handbook of Biological Statistics* 3rd ed. Maryland: Sparky House Publishing, 2014.
- [41] M. Hollander and D. A. Wolfe, "Kendall and Spearman tests," in *Nonparametric statistical methods*, ed New York: John Wiley & Sons, 1973, pp. 185-194.
- [42] ETH_Zürich. "R: Test for Association/Correlation Between Paired Samples". Available: <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/cor.test.html> [Accessed: June 20, 2017]
- [43] A. Stewart, "Basic statistics and epidemiology : a practical guide," in *Basic statistics and epidemiology : a practical guide*, ed Oxford: Radcliffe Pub., 2010, p. 65.

- [44] W. Aigner, A. Rind, and S. Hoffmann, "Comparative Evaluation of an Interactive Time-Series Visualization that Combines Quantitative Data with Qualitative Abstractions," *Computer Graphics Forum*, vol. 31, pp. 995-1004, 2012.
- [45] R. Bade, S. Schlechtweg, and S. Miksch, "Connecting time-oriented data and information to a coherent interactive visualization," presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Vienna, Austria, 2004.
- [46] F. D. Davis, "A technology acceptance model for empirically testing new end-user information systems: Theory and results," Massachusetts Institute of Technology, 1985.
- [47] github/mcaule. "*D3 exploding boxplot by mcaule*" [online]. Available: https://mcaule.github.io/d3_exploding_boxplot/ [Accessed: June 25, 2017]
- [48] T. V. Perneger, "What's wrong with Bonferroni adjustments," *Bmj*, vol. 316, pp. 1236-8, Apr 18 1998.

List of Figures

Figure 1: Classical model of the hematopoiesis in humans	3
Figure 2: Revised model of the pathways of cell differentiation during hematopoiesis	4
Figure 3: first (year 1996) flow cytometry report of the CCRI	8
Figure 4: flow cytometry report of the CCRI	8
Figure 5: Positive vote for the ethics approval for the data analysis of the project for the master thesis	13
Figure 6: Overview of data flow (content of excel files into database)	15
Figure 7: Guideline how to format patient's flow cytometry report to be extracted correctly	17
Figure 8: "patient2db" flow chart.	25
Figure 9: Simplified "parse_xls" flow chart.....	27
Figure 10: Screenshot of regex101.com online regex tester and debugger	28
Figure 11: parse_xls console prompt	28
Figure 12: Main menu (landing page) of work	29
Figure 13: Screenshot of facs2vis.php	32
Figure 14: flow cytometry results visualized in an interactive C3 timeseries chart.	34
Figure 15: Comparison groups for statistical analysis	37
Figure 16: enhanced entity–relationship (EER) model of 'facs_db'	42
Figure 17: Error detection.	44
Figure 18: Snippet of the Excel documentation table	45
Figure 19: Line chart with date input errors	46
Figure 20: Line chart after correction	47
Figure 21: Graft source comparison, 30 (+/- 5) days after HSCT (TX)	49

Figure 22: Graft source comparison, 100 (+/-10) days after HSCT (TX).....	50
Figure 23: Graft source comparison, 180 (+/-20) days after HSCT (TX).....	51
Figure 24: Graft source comparison, 365 (+/-20) days after HSCT (TX).....	52
Figure 25: Donor's age comparison, 30 (+/-5) days after HSCT (TX).....	53
Figure 26: Donor's age comparison, 100 (+/-10) days after HSCT (TX).....	54
Figure 27: Donor's age comparison, 180 (+/-20) days after HSCT (TX).....	55
Figure 28: Donor's age comparison, 365 (+/-20) days after HSCT (TX).....	56
Figure 29: Recipient's age comparison, 30 (+/-5) days after HSCT (TX).....	57
Figure 30: Recipient's age comparison, 100 (+/-10) days after HSCT (TX).....	58
Figure 31: Recipient's age comparison, 180 (+/-20) days after HSCT (TX).....	59
Figure 32: Recipient's age comparison, 365 (+/-20) days after HSCT (TX).....	60
Figure 33: Distribution of data (category graft source comparison)	61
Figure 34: Distribution of data (category donor's age comparison)	62
Figure 35: Distribution of data (category recipient's age comparison)	63
Figure 36: Spearman's rank correlation (category recipient's age),.....	66
Figure 37: Spearman's rank correlation (category donor's age)	67
Figure 38: SPLOM recipient's age, donor's age, naïve CD4 count, day 30 (+/-5) after HSCT.....	68
Figure 39: SPLOM recipient's age, donor's age, naïve CD4 count, day 100 (+/-10) after HSCT	68
Figure 40: SPLOM recipient's age, donor's age, naïve CD4 count, day 180 (+/-20) after HSCT.....	69
Figure 41: SPLOM recipient's age, donor's age, naïve CD4 count, day 365(+/-20) after HSCT	69
Figure 42: Exploding boxplots.....	75

List of Tables

Table 1: allogenic HSCT in Austria, age at HSCT <18, from 2007-2016 [1]	1
Table 2: Required software for the master thesis and the related work	14
Table 3: Table „analysis“ of database facs_db.	21
Table 4: Table „patient“ of database facs_db.“	22
Table 5: Table „transplantation“ of database facs_db.“	22
Table 6: Table „diag_code“ of database facs_db.“	23
Table 7: Table „HSC_Source“ of database facs_db.“	23
Table 8: C3 time series graph color selection	33
Table 9: definition of strength of correlation	37
Table 10: Comparison of groups (graft source)	64
Table 11 Comparison of groups (Donor's age)	65
Table 12: Comparison of groups (Recipient's age)	65

Listings

Listing 1: AJAX method to load returned data into patient dropdown menu..... 30

List of Abbreviations

A	
AJAX	Asynchronous JavaScript and XML
ALL	acute lymphatic leukemia
allo-HSCT	allogenic hematopoietic stem cell transplantation
AML	acute myeloid leukemia
B	
BAL	bronchial alveolar lavage
BM	bone marrow
C	
CB	cord blood
CCRI	St. Anna Children's Research Institute
CD	cluster of differentiation
CML	chronic myeloid leukemia
CMML	chronic myelomonocytic leukemia
CSS	cascading style sheets
D	
DLI	donor lymphocyte cell infusion
E	
EMP	erythro-myeloid progenitor cell
eng.	english
ER	entity relationship
F	
FACS®	fluorescence-activated cell sorting
FISH	fluorescence in situ hybridization
G	
Granulo	granulocytes
GUI	graphical user interface
GvHD	graft versus host disease
H	
HD	Huntington's disease
HLA	human leukocyte antigen
HSC	hematopoietic cell
HSCT	haematopoietic stem cell transplantation
HTML	hypertext markup language

I	
IP	internet protocoll
IT	internet technology
J	
JSON	JavaScript Object Notation
K	
KM	(German) bone marrow
L	
LCH	Langerhans cell histiocytosis
LMPP	lympho-myeloid progenitor cell
M	
MD	medical doctor
MDS	myelodysplastic syndrome
MM	mismatch
Mono	monocyte
MPP	multipotent progenitor cell
MS	Microsoft
N	
n.s.	not significant
NBL	neuroblastoma
NC	nuclated cell
NK	natural killer cells
NKT	natural killer T-cells
NN	placeholder for a number
P	
PBSC	peripheral blood stem cells
PCR	polymerase chain reaction
PHP	Hypertext Preprocessor
R	
RMS	rhabdomyosarcoma
RQ	research question
S	
SAA	severe aplastic anemia
SCID	severe combined imune deficnecy
SEM	standard error of the mean
sig.	significant
SQL	structured query language
SVG	scalable vector graphic
T	
TH	threshold
TX	transplantation (analogous to HSCT)

U	
UPN	unique patient number
μL	micro liter
V	
vis	visualization
W	
WBC	white blood cell
WP	work package
X	
XML	Extensible Markup Language

Appendix

A. Ethic approval



Borschkegasse 8b/6
1090 Wien, Österreich
T +43(0)1 404 00-21470, 22440
F +43(0)1 404 00-16900
ethik-kom@meduniwien.ac.at
<http://ethikkommission.meduniwien.ac.at/>

Votum:

EK Nr: 2231/2016

Projekttitel: FACS2VIS Engraftmentcharakteristik nach allogener Stammzelltransplantation - Retrospektive Datenanalyse basierend auf interaktiver Visualisierung und statistischer Korrelationsrechnung der Leukozytenzellzahlen.

Antragsteller/in: Herr BSc Jakob Winkler

Institution: Fachhochschule St. Pölten

Sponsor: MUW

Teilnehmende Prüfzentren:

Ethik-Kommission	Prüfzentrum	Prüfärztin/arzt
Ethikkommission der Medizinischen Universität Wien	St. Anna Kinderspital - Stammzelltransplantationseinheit - Station 1A	Herr Dr. Herbert Pichler

Die Stellungnahme der Ethik-Kommission erfolgt aufgrund folgender eingereichter Unterlagen:

Lebenslauf (CV)

Name	Version	Datum
Curriculum Vitae HP Version 1.2 unterschrieben	1.2	29.11.2016

Sonstige

Name	Version	Datum
Verpflichtungserklärung_WinklerJ	1.0	01.12.2016
dvrAuszug-ccri	1.0	18.01.2017
Ethikantrag Visualisierung FACS Jakob Winkler 20170118 Version 2_0_AenderungenMarkiert	2.0	18.01.2017



Studienprotokoll (Prüfplan)

Name	Version	Datum
Ethikantrag Visualisierung FACS Jakob Winkler 20161201 Version 1_2	1.2	01.12.2016
Ethikantrag Visualisierung FACS Jakob Winkler 20170118 Version 2_0	2.0	18.01.2017

Die Kommission fasst folgenden Beschluss (mit X markiert):

<input checked="" type="checkbox"/>	Es besteht kein Einwand gegen die Durchführung der Studie.
-------------------------------------	--

Ergänzende Kommentare der Sitzung am 10.01.2017:

Zu Prüfplan und Antrag:

1) Es ist angeführt, dass in der Datenbank die Daten personenbezogen gespeichert werden (Vorname, Nachname).

Für das Speichern von sensiblen, personenbezogenen Daten wären zusätzliche Maßnahmen notwendig wie z.B.

- Meldung beim Datenverarbeitungsregister (DVR)
- Zugriffsprotokollierung lt DSGVO (wer hat wann wie auf die Daten zugegriffen)
- Spezielle Schutzmaßnahmen lt DSGVO

Für eine retrospektive Datenanalyse ist aber eine personenbezogene Speicherung nicht notwendig; eine Pseudonymisierung reicht aus.

2) Bei einem Datenexport für externe Personen/Stellen ist eine Pseudonymisierung notwendig, die keinen Bezug zum Patienten zulässt. Die interne Patienten-ID (HSCT) ist dafür nicht geeignet - hier wäre eine eigene Nummerierung vorzusehen (betrifft auch Punkt 11).

Die Anmerkung "Folglich bestünde somit die Möglichkeit, zusätzliche klinische Daten der PatientInnen für statistische Korrelationen zu verwenden." ist problematisch und sollte entfernt oder geändert werden. Bei Erweiterungen der Fragestellung wäre ein neuer Antrag oder ein Ergänzungsantrag zu stellen. Ein Zusammenhang mit dem Export ist auch nicht gegeben (wirkt so im Text); eine Verknüpfung der Daten durch Externe sollte durch ein Pseudonym ausgeschlossen werden.

Die Ethik-Kommission ersucht die Antragsteller, bei der Wiedervorlage von geänderten Unterlagen ein Exemplar mit hervorgehobenen Änderungen beizulegen.

Ergänzende Kommentare:

Nachtrag vom 23. Jänner 2017:

Die Antragsteller legen am 19.01.2017 überarbeitete Unterlagen vor, die von der Ethik-Kommission akzeptiert werden.



Die aktuelle Mitgliederliste der Ethik-Kommission ist unter folgender Adresse abrufbar:

<http://ethikkommission.meduniwien.ac.at/ethik-kommission/mitglieder/>


Mitglieder der Ethik-Kommission, die für diesen Tagesordnungspunkt als befangen anzusehen waren und daher laut Geschäftsordnung an der Entscheidungsfindung/Abstimmung nicht

teilgenommen haben: Herr Priv. Doz. OA. Dr. Andishe ATTARBASCHI, Frau Univ.Prof. Christina Peters

ACHTUNG: Unter Berücksichtigung der "ICH-Guideline for Good Clinical Practice" gilt dieser Beschluss ein Jahr ab Datum der Ausstellung. Gegebenenfalls hat der Antragsteller eine Verlängerung der Gültigkeit rechtzeitig zu beantragen.

Dieses Dokument ist für berechnigte Benutzer/innen in digitaler Form unter folgender Adresse abrufbar:

<https://ekmeduniwien.at/vote/10950/download/>

	Unterzeichner	Dr. Martin Bernhard Brunner
	Datum/Zeit-UTC	2017-01-26T16:29:12Z
	Prüfinformation	Informationen zur Prüfung der elektronischen Signatur finden Sie unter: https://www.signaturpruefung.gv.at



B. Declaration of consent clinical data export

Dr. Herbert Pichler
Univ. Doz. Dr^a. Susanne Matthes - Leodolter
Stammzelltransplantationseinheit, 1A
St. Anna Kinderspital
Kinderspitalg.6
1090 Wien

Wien, 18.01.2017

Datenverwendung im Rahmen der Masterarbeit Projekt „FACS2VIS“

Hiermit wird Herrn Jakob Winkler, geb 30.11.1984 bestätigt, dass er im Rahmen seiner Masterarbeit (12/2016-08/2017) in der St. Anna Kinderkrebsforschung zur Erstellung der Datenbank FACS2VIS Daten aus den Dokumentationstabellen stamm1c.xls, DLINACHK.xls des St. Anna Kinderspitals verwenden darf. Die Daten dürfen nicht an Dritte weitergegeben werden und sind in einer zugriffsgeschützten Datenbank zu speichern.

Auswertungsergebnisse dürfen in der Masterarbeit nur so präsentiert werden, dass sie keinen direkten Rückschluss auf das Individuum zulassen.

Folgende Daten dürfen ausschließlich für dieses Projekt verwendet werden:

- TX (chronologische Transplantationsnummer)
- TR_NR (Zahl der Transplantationen pro PatientIn)
- FAMILIENNAME (Familiennamen des/der PatientIn)
- VOR_NAME (Vorname des/der PatientIn)
- GEB_DAT (PatientIn Geburtsdatum)
- DIAG_CODE + Kommentar (Diagnosecode der Erkrankung)
- ERST_DIAG_TXT (Erkrankung)
- KMT_DAT (Datum der HSCT)
- KG_KMT (PatientIn Gewicht bei HSCT)
- KMT_ART (Art der Transplantation [autolog/allogen])
- SZ-ART (Art des Grafts [BM/PBSC/CB])
- S_GE (SpenderIn Geschlecht)
- S-GRAD (Verwandschaftsverhältnis Spender / Empfänger)
- S_GEBDAT (Geburtsdatum des Spenders)
- CD34kg (Anzahl der CD34+ Zellen/kg KG im Graft)
- TcellKG (Anzahl der CD3+ Zellen/kg KG im Graft)
- ZAHL MM (HLA Mismatch Donor/Recipient)
- Patienten, die DLI erhalten haben (aus DLINACHK.XLS)


Dr. Herbert Pichler


Univ. Doz. Dr^a. Susanne Matthes - Leodolter

C. Color panel for flow cytometry

SOP CP-AN-004 MD1-1D

Seite 1 von 1

Cocktails für Routine Diagnostik ab 17.07.2015

15µl für 0,15 x 10E⁹ Leukos

Violett I		AX488	PE	PE-CF594	PerCP	PerCP e-FL710	PE-Cy7	APC	Alexa Fl 700	APC-Cy7	PBS+ Na-Acid
1. Färbung											
Syto 41		15	33	45RA	45	3	10	34	38	19+20	
Einzelz zugeben		50µl	200µl	1µl	200µl	50µl	25µl	50µl	50µl	25µl+25µl	325µl
1-5µl											
2. Färbung											
Syto41	BV510	FITC	PE	PE-CF594	PerCP	PerCP e-FL710	PE-Cy7	APC		APC-Cy7	
Einzelz zugeben	19	ab	gd	3	45	4	56	16		8	
1-5µl	50µl	200µl	200µl	2,5µl	200µl	50µl	50µl	25µl		50µl	175µl
3. Färbung											
Syto 41		FITC	PE	PE-CF594	PerCP	PerCP e-FL710	PE-Cy7	APC		APC-Cy7	
Einzelz zugeben		27	31	45RA	45	4	56	3		8	
1-5µl		200µl	200µl	1µl	200µl	50µl	50µl	50µl		50µl	200µl
4. Färbung											
Syto41		AX488	PE	PE-CF594	PerCP	PerCP e-FL710	PE-Cy7	APC	Alexa Fl 700	APC-Cy7	
Einzelz zugeben		15	34	45RA	45	33	10	71	38	19+20	
1-5µl		50µl	200µl	1µl	200µl	100µl	25µl	50µl	50µl	25µl+25µl	275µl
Vor Transplant		AX488	PE		PerCP			APC		APC-Cy7	
		FITC									
		15+45	56		3			19		14	
		15µl+30µl	30µl		60µl			15µl		15µl	135µl
Nach Transplant		AX488	PE		PerCP		PE-Cy7	APC		APC-Cy7	
Nippl		FITC									
		15+45	38		3		56	34		14	
		15µl+30µl	60µl		60µl		15µl	15µl		15µl	60µl
ECP		AX488	PE		PerCP			APC			
(+/- TruCount Tube)		15	33		45			34			
T-Zell Subtypen (Cocktail für 15 Färbungen)											
HLA-DR auf CD3			PE		PerCP		PE-Cy7	APC		APC-Cy7	
			HLA-DR		3		33	19		45	
			60µl		60µl		15µl	15µl		15µl	135µl
CD25 auf naiven/ memory CD4		FITC		PE-CF594	PerCP		PE-Cy7	APC		APC-Cy7	
		45RO		45RA	3		25	4		45	
		60µl		1µl	60µl		7,5µl	30µl		30µl	60µl

D. R Script for explorative statistical analysis

```
# Exprotative Statistical Analysis for
# Master Thesis
# Jakob Winkler
# 04.07.2017

# Comparison Graft Source

PBSC_depl_day30 <-
read.table("F:/ajax/exportFACS2VIS_materialPBSC_depl_day30.csv",
header=T, sep=";")
PBSC_depl_day100 <-
read.table("F:/ajax/exportFACS2VIS_materialPBSC_depl_day100.csv",
header=T, sep=";")
PBSC_depl_day180 <-
read.table("F:/ajax/exportFACS2VIS_materialPBSC_depl_day180.csv",
header=T, sep=";")
PBSC_depl_day365 <-
read.table("F:/ajax/exportFACS2VIS_materialPBSC_depl_day365.csv",
header=T, sep=";")

PBSC_day30 <-
read.table("F:/ajax/exportFACS2VIS_materialPBSC_day30.csv", header=T,
sep=";")
PBSC_day100 <-
read.table("F:/ajax/exportFACS2VIS_materialPBSC_day100.csv", header=T,
sep=";")
PBSC_day180 <-
read.table("F:/ajax/exportFACS2VIS_materialPBSC_day180.csv", header=T,
sep=";")
PBSC_day365 <-
read.table("F:/ajax/exportFACS2VIS_materialPBSC_day365.csv", header=T,
sep=";")

BM_day30 <- read.table("F:/ajax/exportFACS2VIS_materialBM_day30.csv",
header=T, sep=";")
BM_day100 <-
read.table("F:/ajax/exportFACS2VIS_materialBM_day100.csv", header=T,
sep=";")
BM_day180 <-
read.table("F:/ajax/exportFACS2VIS_materialBM_day180.csv", header=T,
sep=";")
BM_day365 <-
read.table("F:/ajax/exportFACS2VIS_materialBM_day365.csv", header=T,
sep=";")

# -----

# Comparison Sex

donor_male_day30 <-
read.table("F:/ajax/exportFACS2VIS_DonorSex_male_day30.csv", header=T,
sep=";")
donor_male_day100 <-
read.table("F:/ajax/exportFACS2VIS_DonorSex_male_day100.csv",
header=T, sep=";")
donor_male_day180 <-
read.table("F:/ajax/exportFACS2VIS_DonorSex_male_day180.csv",
header=T, sep=";")
```

```

donor_male_day365 <-
read.table("F:/ajax/exportFACS2VIS_DonorSex_male_day365.csv",
header=T, sep=";")

donor_female_day30 <-
read.table("F:/ajax/exportFACS2VIS_DonorSex_female_day30.csv",
header=T, sep=";")
donor_female_day100 <-
read.table("F:/ajax/exportFACS2VIS_DonorSex_female_day100.csv",
header=T, sep=";")
donor_female_day180 <-
read.table("F:/ajax/exportFACS2VIS_DonorSex_female_day180.csv",
header=T, sep=";")
donor_female_day365 <-
read.table("F:/ajax/exportFACS2VIS_DonorSex_female_day365.csv",
header=T, sep=";")

# -----

# Comparison Recipient Age

recipientage10_day30 <-
read.table("F:/ajax/exportFACS2VIS_RecipientAge10_day30.csv",
header=T, sep=";")
recipientage10_day100 <-
read.table("F:/ajax/exportFACS2VIS_RecipientAge10_day100.csv",
header=T, sep=";")
recipientage10_day180 <-
read.table("F:/ajax/exportFACS2VIS_RecipientAge10_day180.csv",
header=T, sep=";")
recipientage10_day365 <-
read.table("F:/ajax/exportFACS2VIS_RecipientAge10_day365.csv",
header=T, sep=";")

recipientage10_day30 <-
read.table("F:/ajax/exportFACS2VIS_RecipientAge10_day30.csv",
header=T, sep=";")
recipientage10_day100 <-
read.table("F:/ajax/exportFACS2VIS_RecipientAge10_day100.csv",
header=T, sep=";")
recipientage10_day180 <-
read.table("F:/ajax/exportFACS2VIS_RecipientAge10_day180.csv",
header=T, sep=";")
recipientage10_day365 <-
read.table("F:/ajax/exportFACS2VIS_RecipientAge10_day365.csv",
header=T, sep=";")

# -----

# Comparison Donor Age (3 Groups)

donorage10_day30 <-
read.table("F:/ajax/exportFACS2VIS_donorAge10_day30.csv", header=T,
sep=";")
donorage10_day100 <-
read.table("F:/ajax/exportFACS2VIS_donorAge10_day100.csv", header=T,
sep=";")
donorage10_day180 <-
read.table("F:/ajax/exportFACS2VIS_donorAge10_day180.csv", header=T,
sep=";")
donorage10_day365 <-
read.table("F:/ajax/exportFACS2VIS_donorAge10_day365.csv", header=T,
sep=";")

donorage1020_day30 <- read.table("F:/ajax/exportFACS2VIS_donorAge10-
20_day30.csv", header=T, sep=";")

```

```

donorage1020_day100 <- read.table("F:/ajax/exportFACS2VIS_donorAge10-
20_day100.csv", header=T, sep=";")
donorage1020_day180 <- read.table("F:/ajax/exportFACS2VIS_donorAge10-
20_day180.csv", header=T, sep=";")
donorage1020_day365 <- read.table("F:/ajax/exportFACS2VIS_donorAge10-
20_day365.csv", header=T, sep=";")

donorage20_day30 <-
read.table("F:/ajax/exportFACS2VIS_donorAge20_day30.csv", header=T,
sep=";")
donorage20_day100 <-
read.table("F:/ajax/exportFACS2VIS_donorAge20_day100.csv", header=T,
sep=";")
donorage20_day180 <-
read.table("F:/ajax/exportFACS2VIS_donorAge20_day180.csv", header=T,
sep=";")
donorage20_day365 <-
read.table("F:/ajax/exportFACS2VIS_donorAge20_day365.csv", header=T,
sep=";")

#-----
#-----

# Spalten dazu

# der t-test oder kruskal wallies - test verlangen nach numerische
variablen, deshalb hier die Stammzellquelle numerisch codiert;
PBSC_depl_day30$sourceN <- c(3)
PBSC_depl_day100$sourceN <- c(3)
PBSC_depl_day180$sourceN <- c(3)
PBSC_depl_day365$sourceN <- c(3)
PBSC_day30$sourceN <- c(2)
PBSC_day100$sourceN <- c(2)
PBSC_day180$sourceN <- c(2)
PBSC_day365$sourceN <- c(2)
BM_day30$sourceN <- c(1)
BM_day100$sourceN <- c(1)
BM_day180$sourceN <- c(1)
BM_day365$sourceN <- c(1)

# donor sex numerisch codiert
donor_male_day30$donorsexN <- c(1)
donor_male_day100$donorsexN <- c(1)
donor_male_day180$donorsexN <- c(1)
donor_male_day365$donorsexN <- c(1)
donor_female_day30$donorsexN <- c(2)
donor_female_day100$donorsexN <- c(2)
donor_female_day180$donorsexN <- c(2)
donor_female_day365$donorsexN <- c(2)

# Recipient Age numerisch codiert

recipientageu10_day30$recipientAgeN <- c(1)
recipientageu10_day100$recipientAgeN <- c(1)
recipientageu10_day180$recipientAgeN <- c(1)
recipientageu10_day365$recipientAgeN <- c(1)
recipientage10_day30$recipientAgeN <- c(2)
recipientage10_day100$recipientAgeN <- c(2)
recipientage10_day180$recipientAgeN <- c(2)

```

```

recipientage10_day365$recipientAgeN <- c(2)

# Donor Age (3 Groups) numerisch kodiert

donorage10_day30$donorageN <- c(1)
donorage10_day100$donorageN <- c(1)
donorage10_day180$donorageN <- c(1)
donorage10_day365$donorageN <- c(1)
donorage1020_day30$donorageN <- c(2)
donorage1020_day100$donorageN <- c(2)
donorage1020_day180$donorageN <- c(2)
donorage1020_day365$donorageN <- c(2)
donorage20_day30$donorageN <- c(3)
donorage20_day100$donorageN <- c(3)
donorage20_day180$donorageN <- c(3)
donorage20_day365$donorageN <- c(3)

#-----
----

# für ANOVA - Faktor - Variable notwendig;
PBSC_depl_day30$source <- c("PBSC_depl")
PBSC_depl_day100$source <- c("PBSC_depl")
PBSC_depl_day180$source <- c("PBSC_depl")
PBSC_depl_day365$source <- c("PBSC_depl")
PBSC_day30$source <- c("PBSC")
PBSC_day100$source <- c("PBSC")
PBSC_day180$source <- c("PBSC")
PBSC_day365$source <- c("PBSC")
BM_day30$source <- c("BM")
BM_day100$source <- c("BM")
BM_day180$source <- c("BM")
BM_day365$source <- c("BM")

# donor sex als Faktor
donor_male_day30$donorsex <- c("male")
donor_male_day100$donorsex <- c("male")
donor_male_day180$donorsex <- c("male")
donor_male_day365$donorsex <- c("male")
donor_female_day30$donorsex <- c("female")
donor_female_day100$donorsex <- c("female")
donor_female_day180$donorsex <- c("female")
donor_female_day365$donorsex <- c("female")

# recipient age als Faktor

recipientageu10_day30$recipientAge <- c("u10")
recipientageu10_day100$recipientAge <- c("u10")
recipientageu10_day180$recipientAge <- c("u10")
recipientageu10_day365$recipientAge <- c("u10")
recipientage10_day30$recipientAge <- c("10+")
recipientage10_day100$recipientAge <- c("10+")
recipientage10_day180$recipientAge <- c("10+")
recipientage10_day365$recipientAge <- c("10+")

```

```

# Donor Age als Faktor

donorage10_day30$donorage <- c("u10")
donorage10_day100$donorage <- c("u10")
donorage10_day180$donorage <- c("u10")
donorage10_day365$donorage <- c("u10")
donorage1020_day30$donorage <- c("10-20")
donorage1020_day100$donorage <- c("10-20")
donorage1020_day180$donorage <- c("10-20")
donorage1020_day365$donorage <- c("10-20")
donorage20_day30$donorage <- c("20+")
donorage20_day100$donorage <- c("20+")
donorage20_day180$donorage <- c("20+")
donorage20_day365$donorage <- c("20+")

#-----

# die 4 files aneinander gehängt
source_day30 <- rbind(BM_day30,PBSC_day30,PBSC_depl_day30)
source_day100 <- rbind(BM_day100,PBSC_day100,PBSC_depl_day100)
source_day180 <- rbind(BM_day180,PBSC_day180,PBSC_depl_day180)
source_day365 <- rbind(BM_day365,PBSC_day365,PBSC_depl_day365)

donorsex_day30 <- rbind(donor_male_day30,donor_female_day30)
donorsex_day100 <- rbind(donor_male_day100,donor_female_day100)
donorsex_day180 <- rbind(donor_male_day180,donor_female_day180)
donorsex_day365 <- rbind(donor_male_day365,donor_female_day365)

recipientage_day30 <- rbind(recipientageu10_day30,
recipientage10_day30)
recipientage_day100 <- rbind(recipientageu10_day100,
recipientage10_day100)
recipientage_day180 <- rbind(recipientageu10_day180,
recipientage10_day180)
recipientage_day365 <- rbind(recipientageu10_day365,
recipientage10_day365)

donorage_day30 <- rbind(donorage10_day30, donorage1020_day30,
donorage20_day30)
donorage_day100 <- rbind(donorage10_day100, donorage1020_day100,
donorage20_day100)
donorage_day180 <- rbind(donorage10_day180, donorage1020_day180,
donorage20_day180)
donorage_day365 <- rbind(donorage10_day365, donorage1020_day365,
donorage20_day365)

#-----

#STATISTISCHE TESTS

#-----

# Stammzellquelle / Graft Source
#-----

```

```

# 3 Gruppen
# Der Kruskal-Wallis-Test (nach William Kruskal und Wilson Allen
Wallis; auch H-Test) ist ein parameterfreier statistischer Test,
# mit dem im Rahmen einer Varianzanalyse getestet wird, ob unabhängige
Stichproben (Gruppen oder Messreihen) hinsichtlich
# einer ordinalskalierten Variable einer gemeinsamen Population
entstammen.
# Er ähnelt einem Mann-Whitney-U-Test und basiert wie dieser auf
Rangplatzsummen, mit dem Unterschied,
# dass er für den Vergleich von mehr als zwei Gruppen angewendet
werden kann.
# Da multiples Testen (4 Messzeitpunkte) signifikant bei  $p < 0,0125$ 
(Bonferroni Korrektur des Alpha fehlers)
# Keine post-hoc pairwise test möglich, da overall p nicht  $< 0,0125$ 

kruskal.test(CD4_naiv_ABS~sourceN, data=source_day30)
#pairwise.wilcox.test(source_day30$CD4_naiv_ABS, source_day30$source,
p.adj = "bonferroni", paired = FALSE)
pairwise.wilcox.test(source_day30$CD4_naiv_ABS, source_day30$source,
paired = FALSE)

kruskal.test(CD4_naiv_ABS~sourceN, data=source_day100)
#pairwise.wilcox.test(source_day100$CD4_naiv_ABS,
source_day100$source, p.adj = "bonferroni", paired = FALSE)
pairwise.wilcox.test(source_day100$CD4_naiv_ABS, source_day100$source,
paired = FALSE)

kruskal.test(CD4_naiv_ABS~sourceN, data=source_day180)
#pairwise.wilcox.test(source_day180$CD4_naiv_ABS,
source_day180$source, p.adj = "bonferroni", paired = FALSE)
pairwise.wilcox.test(source_day180$CD4_naiv_ABS, source_day180$source,
paired = FALSE)

kruskal.test(CD4_naiv_ABS~sourceN, data=source_day365)
#pairwise.wilcox.test(source_day365$CD4_naiv_ABS,
source_day365$source, p.adj = "bonferroni", paired = FALSE)
pairwise.wilcox.test(source_day365$CD4_naiv_ABS, source_day365$source,
paired = FALSE)

plot_source_hist<-par(mfrow=c(2,2))
hist30_graft<-hist(source_day30$CD4_naiv_ABS, main="naive CD4 [day 30
after HSCT] Graft.Source", xlab="GS/naive CD4 cells [/µl]")
hist100_graft<-hist(source_day100$CD4_naiv_ABS, main="naive CD4 [day
100 after HSCT] Graft.Source", xlab="GS/naive CD4 cells [/µl]")
hist180_graft<-hist(source_day180$CD4_naiv_ABS, main="naive CD4 [day
180 after HSCT] Graft.Source", xlab="GS/naive CD4 cells [/µl]")
hist365_graft<-hist(source_day365$CD4_naiv_ABS, main="naive CD4 [day
365 after HSCT] Graft.Source", xlab="GS/naive CD4 cells [/µl]")
mtext("Histogram DONOR AGE", outer = TRUE, cex = 1.5)
plot_source_hist
x11()

hist30_graft

#-----
# DONOR AGE
#-----

```

```

# 3 Gruppen
# Der Kruskal-Wallis-Test (nach William Kruskal und Wilson Allen
Wallis; auch H-Test) ist ein parameterfreier statistischer Test,
# mit dem im Rahmen einer Varianzanalyse getestet wird, ob unabhängige
Stichproben (Gruppen oder Messreihen) hinsichtlich
# einer ordinalskalierten Variable einer gemeinsamen Population
entstammen.[1] Er ähnelt einem Mann-Whitney-U-Test und basiert wie
dieser auf Rangplatzsummen, mit dem Unterschied, dass er für den
Vergleich von mehr als zwei Gruppen angewendet werden kann.
# Da multiples Testen (4 Messzeitpunkte) signifikant bei  $p < 0,0125$ 
(Bonferroni Korrektur des Alpha fehlers)
# Keine post-hoc pairwise test möglich, wenn overall p nicht  $< 0,0125$ 

kruskal.test(CD4_naiv_ABS~donorageN, data=donorage_day30)
pairwise.wilcox.test(donorage_day30$CD4_naiv_ABS,
donorage_day30$donorage, p.adjust.method = "none", paired = FALSE)
#pairwise.wilcox.test(donorage_day30$CD4_naiv_ABS,
donorage_day30$donorage, paired = FALSE)
kruskal.test(CD4_naiv_ABS~donorageN, data=donorage_day100)
pairwise.wilcox.test(donorage_day100$CD4_naiv_ABS,
donorage_day100$donorage, p.adjust.method = "none", paired = FALSE)
#pairwise.wilcox.test(donorage_day100$CD4_naiv_ABS,
donorage_day100$donorage, paired = FALSE)
kruskal.test(CD4_naiv_ABS~donorageN, data=donorage_day180)
pairwise.wilcox.test(donorage_day180$CD4_naiv_ABS,
donorage_day180$donorage, p.adjust.method = "none", paired = FALSE)
#pairwise.wilcox.test(donorage_day180$CD4_naiv_ABS,
donorage_day180$donorage, paired = FALSE)
kruskal.test(CD4_naiv_ABS~donorageN, data=donorage_day365)
pairwise.wilcox.test(donorage_day365$CD4_naiv_ABS,
donorage_day365$donorage, p.adjust.method = "none", paired = FALSE)
#pairwise.wilcox.test(donorage_day365$CD4_naiv_ABS,
donorage_day365$donorage, paired = FALSE)

#Correlation
#-----

donorage_day30$date_diff_rec <- as.Date(donorage_day30$dateOfTx) -
as.Date(donorage_day30$pat_birthday)
donorage_day30$date_diff_rec_years=round(as.numeric(donorage_day30$dat
e_diff_rec/365), digits =1)
donorage_day100$date_diff_rec<- as.Date(donorage_day100$dateOfTx) -
as.Date(donorage_day100$pat_birthday)
donorage_day100$date_diff_rec_years=round(as.numeric(donorage_day100$d
ate_diff_rec/365), digits =1)
donorage_day180$date_diff_rec<- as.Date(donorage_day180$dateOfTx) -
as.Date(donorage_day180$pat_birthday)
donorage_day180$date_diff_rec_years=round(as.numeric(donorage_day180$d
ate_diff_rec/365), digits =1)
donorage_day365$date_diff_rec<- as.Date(donorage_day365$dateOfTx) -
as.Date(donorage_day365$pat_birthday)
donorage_day365$date_diff_rec_years=round(as.numeric(donorage_day365$d
ate_diff_rec/365), digits =1)

donorage_day30$date_diff <- as.Date(donorage_day30$dateOfTx) -
as.Date(donorage_day30$tx_donor_birthday)
donorage_day30$date_diff_donor_years=round(as.numeric(donorage_day30$d
ate_diff/365), digits =1)
correlation<-cor.test(donorage_day30$date_diff_donor_years,
donorage_day30$CD4_naiv_ABS , method="spearman")

```



```

correlation
length(donorage_day30$date_diff)

donorage_day100$date_diff <- as.Date(donorage_day100$dateOfTx)-
as.Date(donorage_day100$tx_donor_birthday)
donorage_day100$date_diff_donor_years=round(as.numeric(donorage_day100
$date_diff/365), digits =1)
correlation<-cor.test(donorage_day100$date_diff_donor_years,
donorage_day100$CD4_naiv_ABS , method="spearman")
correlation
length(donorage_day100$date_diff)

donorage_day180$date_diff <- as.Date(donorage_day180$dateOfTx)-
as.Date(donorage_day180$tx_donor_birthday)
donorage_day180$date_diff_donor_years=round(as.numeric(donorage_day180
$date_diff/365), digits =1)
correlation<-cor.test(donorage_day180$date_diff_donor_years,
donorage_day180$CD4_naiv_ABS , method="spearman")
correlation
length(donorage_day180$date_diff)

donorage_day365$date_diff <- as.Date(donorage_day365$dateOfTx)-
as.Date(donorage_day365$tx_donor_birthday)
donorage_day365$date_diff_donor_years=round(as.numeric(donorage_day365
$date_diff/365), digits =1)
correlation<-cor.test(donorage_day365$date_diff_donor_years,
donorage_day365$CD4_naiv_ABS , method="spearman")
correlation
length(donorage_day365$date_diff)

plot(donorage_day30[,c(12,20,22)],main="Scatterplot naive CD4 [Day 30
after HSCT]")
x11()
plot(donorage_day100[,c(12,20,22)],main="Scatterplot naive CD4 [Day
100 after HSCT]")
x11()
plot(donorage_day180[,c(12,20,22)],main="Scatterplot naive CD4 [Day
180 after HSCT]")
x11()
plot(donorage_day365[,c(12,20,22)],main="Scatterplot naive CD4 [Day
365 after HSCT]")
x11()

#Correlation between donor and recipient

cor.test(donorage_day30[,c(20)], donorage_day30[,c(22)] ,
method="spearman")
cor.test(donorage_day100[,c(20)], donorage_day100[,c(22)] ,
method="spearman")
cor.test(donorage_day180[,c(20)], donorage_day180[,c(22)] ,
method="spearman")
cor.test(donorage_day365[,c(20)], donorage_day365[,c(22)] ,
method="spearman")

# 2x2 Scatterplot
plot_donortage<-par(mfrow=c(2,2))
plot(donorage_day30$date_diff_donor_years,
donorage_day30$CD4_naiv_ABS, main="Scatterplot naive CD4 / Donor's
age [Day 30 after HSCT]",

```

```

    ylab="naive CD4 [/μl] ", xlab="Donor's age at day of HSCT [years]
", pch=19)
plot(donorage_day100$date_diff_donor_years,
donorage_day100$CD4_naiv_ABS, main="Scatterplot naive CD4 / Donor's
age [Day 100 after HSCT]",
    ylab="naive CD4 [/μl] ", xlab="Donor's age at day of HSCT [years]
", pch=19)
plot(donorage_day180$date_diff_donor_years,
donorage_day180$CD4_naiv_ABS, main="Scatterplot naive CD4 / Donor's
age [Day 180 after HSCT]",
    ylab="naive CD4 [/μl] ", xlab="Donor's age at day of HSCT [years]
", pch=19)
plot(donorage_day365$date_diff_donor_years,
donorage_day365$CD4_naiv_ABS, main="Scatterplot naive CD4 / Donor's
age [Day 365 after HSCT]",
    ylab="naive CD4 [/μl] ", xlab="Donor's age at day of HSCT [years]
", pch=19)
plot_donortage
x11()

# 2x2 Scatterplot
plot_donortage<-par(mfrow=c(2,2))
plot(donorage_day30$date_diff_years, donorage_day30$CD4_naiv_ABS,
main="Scatterplot naive CD4 / Donor's age [Day 30 after HSCT]",
    ylab="naive CD4 [/μl] ", xlab="Donor's age at day of HSCT [years]
", pch=19)
plot(donorage_day100$date_diff_years, donorage_day100$CD4_naiv_ABS,
main="Scatterplot naive CD4 / Donor's age [Day 100 after HSCT]",
    ylab="naive CD4 [/μl] ", xlab="Donor's age at day of HSCT [years]
", pch=19)
plot(donorage_day180$date_diff_years, donorage_day180$CD4_naiv_ABS,
main="Scatterplot naive CD4 / Donor's age [Day 180 after HSCT]",
    ylab="naive CD4 [/μl] ", xlab="Donor's age at day of HSCT [years]
", pch=19)
plot(donorage_day365$date_diff_years, donorage_day365$CD4_naiv_ABS,
main="Scatterplot naive CD4 / Donor's age [Day 365 after HSCT]",
    ylab="naive CD4 [/μl] ", xlab="Donor's age at day of HSCT [years]
", pch=19)
plot_donortage
x11()

#--- Normal distribution?
plot_donortage_hist<-par(mfrow=c(2,2))
hist(donorage_day30$CD4_naiv_ABS, main="naive CD4 [day 30 after HSCT]
Donor.Age", xlab="D/naive CD4 cells [/μl]")
hist(donorage_day100$CD4_naiv_ABS, main="naive CD4 [day 100 after
HSCT] Donor.Age", xlab="D/naive CD4 cells [/μl]")
hist(donorage_day180$CD4_naiv_ABS, main="naive CD4 [day 180 after
HSCT] Donor.Age", xlab="D/naive CD4 cells [/μl]")
hist(donorage_day365$CD4_naiv_ABS, main="naive CD4 [day 365 after
HSCT] Donor.Age", xlab="D/naive CD4 cells [/μl]")
plot_donortage_hist
x11()

donorage_day30

#-----
#Donor SEX

```

```

#-----
#2 Gruppen
#Wilcoxon-Mann-Whitney rank sum test (or test U) für 2 Gruppen

wilcox.test(CD4_naiv_ABS ~ donorsex, data=donorsex_day30)
wilcox.test(CD4_naiv_ABS ~ donorsex, data=donorsex_day100)
wilcox.test(CD4_naiv_ABS ~ donorsex, data=donorsex_day180)
wilcox.test(CD4_naiv_ABS ~ donorsex, data=donorsex_day365)

#-----
#Recipient Age
#-----
#2 Gruppen
#nicht korrektes Verfahren: t-test, weil nicht normalverteilt.
#conf.intervall (1-alpha) auf 0,9875 gesteigert, da multiples testen.
Methode: Bonferroni Korrektur

wilcox.test(CD4_naiv_ABS ~ recipientAge , data=recipientage_day30,
conf.level = 0.9875)
wilcox.test(CD4_naiv_ABS ~ recipientAge , data=recipientage_day100,
conf.level = 0.9875)
wilcox.test(CD4_naiv_ABS ~ recipientAge , data=recipientage_day180,
conf.level = 0.9875)
wilcox.test(CD4_naiv_ABS ~ recipientAge , data=recipientage_day365,
conf.level = 0.9875)
wilcox.test(CD4_naiv_ABS ~ recipientAge , data=recipientage_day30,
conf.level = 0.95)
wilcox.test(CD4_naiv_ABS ~ recipientAge , data=recipientage_day100,
conf.level = 0.95)
wilcox.test(CD4_naiv_ABS ~ recipientAge , data=recipientage_day180,
conf.level = 0.95)
wilcox.test(CD4_naiv_ABS ~ recipientAge , data=recipientage_day365,
conf.level = 0.95)

#Correlation
#-----
recipientage_day30$date_diff <- as.Date(recipientage_day30$dateOfTx)-
as.Date(recipientage_day30$pat_birthday)
recipientage_day30$date_diff_years=round(as.numeric(recipientage_day30
$date_diff/365), digits =1)
correlation<-cor.test(recipientage_day30$date_diff_years,
recipientage_day30$CD4_naiv_ABS , method="spearman")
correlation
length(recipientage_day30$date_diff_years)

recipientage_day100$date_diff <-
as.Date(recipientage_day100$dateOfTx)-
as.Date(recipientage_day100$pat_birthday)
recipientage_day100$date_diff_years=round(as.numeric(recipientage_day1
00$date_diff/365), digits =1)
correlation<-cor.test(recipientage_day100$date_diff_years,
recipientage_day100$CD4_naiv_ABS , method="spearman")
correlation
length(recipientage_day100$date_diff_years)

```

```

recipientage_day180$date_diff <-
as.Date(recipientage_day180$dateOfTx)-
as.Date(recipientage_day180$pat_birthday)
recipientage_day180$date_diff_years=round(as.numeric(recipientage_day1
80$date_diff/365), digits =1)
correlation<-cor.test(recipientage_day180$date_diff_years,
recipientage_day180$CD4_naiv_ABS , method="spearman")
correlation
length(recipientage_day180$date_diff_years)

recipientage_day365$date_diff <-
as.Date(recipientage_day365$dateOfTx)-
as.Date(recipientage_day365$pat_birthday)
recipientage_day365$date_diff_years=round(as.numeric(recipientage_day3
65$date_diff/365), digits =1)
correlation<-cor.test(recipientage_day365$date_diff_years,
recipientage_day365$CD4_naiv_ABS , method="spearman")
correlation
length(recipientage_day365$date_diff_years)

# 2x2 Scatterplot
plot_recipientage<-par(mfrow=c(2,2))
plot(recipientage_day30$date_diff_years,
recipientage_day30$CD4_naiv_ABS, main="Scatterplot naive CD4 /
Recipient's age [Day 30 after HSCT]",
      ylab="naive CD4 [/µl] ", xlab="Recipient's age at day of HSCT
[years] ", pch=19)
plot(recipientage_day100$date_diff_years,
recipientage_day100$CD4_naiv_ABS, main="Scatterplot naive CD4 /
Recipient's age [Day 100 after HSCT]",
      ylab="naive CD4 [/µl] ", xlab="Recipient's age at day of HSCT
[years] ", pch=19)
plot(recipientage_day180$date_diff_years,
recipientage_day180$CD4_naiv_ABS, main="Scatterplot naive CD4 /
Recipient's age [Day 180 after HSCT]",
      ylab="naive CD4 [/µl] ", xlab="Recipient's age at day of HSCT
[years] ", pch=19)
plot(recipientage_day365$date_diff_years,
recipientage_day365$CD4_naiv_ABS, main="Scatterplot naive CD4 /
Recipient's age [Day 365 after HSCT]",
      ylab="naive CD4 [/µl] ", xlab="Recipient's age at day of HSCT
[years] ", pch=19)
plot_recipientage

plot_recipientage_hist<-par(mfrow=c(2,2))
hist(recipientage_day30$CD4_naiv_ABS, main="naive CD4 [day 30 after
HSCT] Recipient.Age", xlab="R/naive CD4 cells [/µl]")
hist(recipientage_day100$CD4_naiv_ABS, main="naive CD4 [day 100 after
HSCT] Recipient.Age", xlab="R/naive CD4 cells [/µl]")
hist(recipientage_day180$CD4_naiv_ABS, main="naive CD4 [day 180 after
HSCT] Recipient.Age", xlab="R/naive CD4 cells [/µl]")
hist(recipientage_day365$CD4_naiv_ABS, main="naive CD4 [day 365 after
HSCT] Recipient.Age", xlab="R/naive CD4 cells [/µl]")
mtext("Histogram RECIPIENT AGE", outer = TRUE, cex = 1.5)
plot_recipientage_hist

```

A. Content of attached digital medium

The attached disc contains the following components

- database design
- declaration of consent (HSCT ward St. Anna Children Hospital, Vienna)
- development (programs for data scraping and web-based visualization)
- ethics approval
- images
- literature
- pres_video (screen capture of web-based visualization, interactive line chart)
- this thesis in .pdf format.

