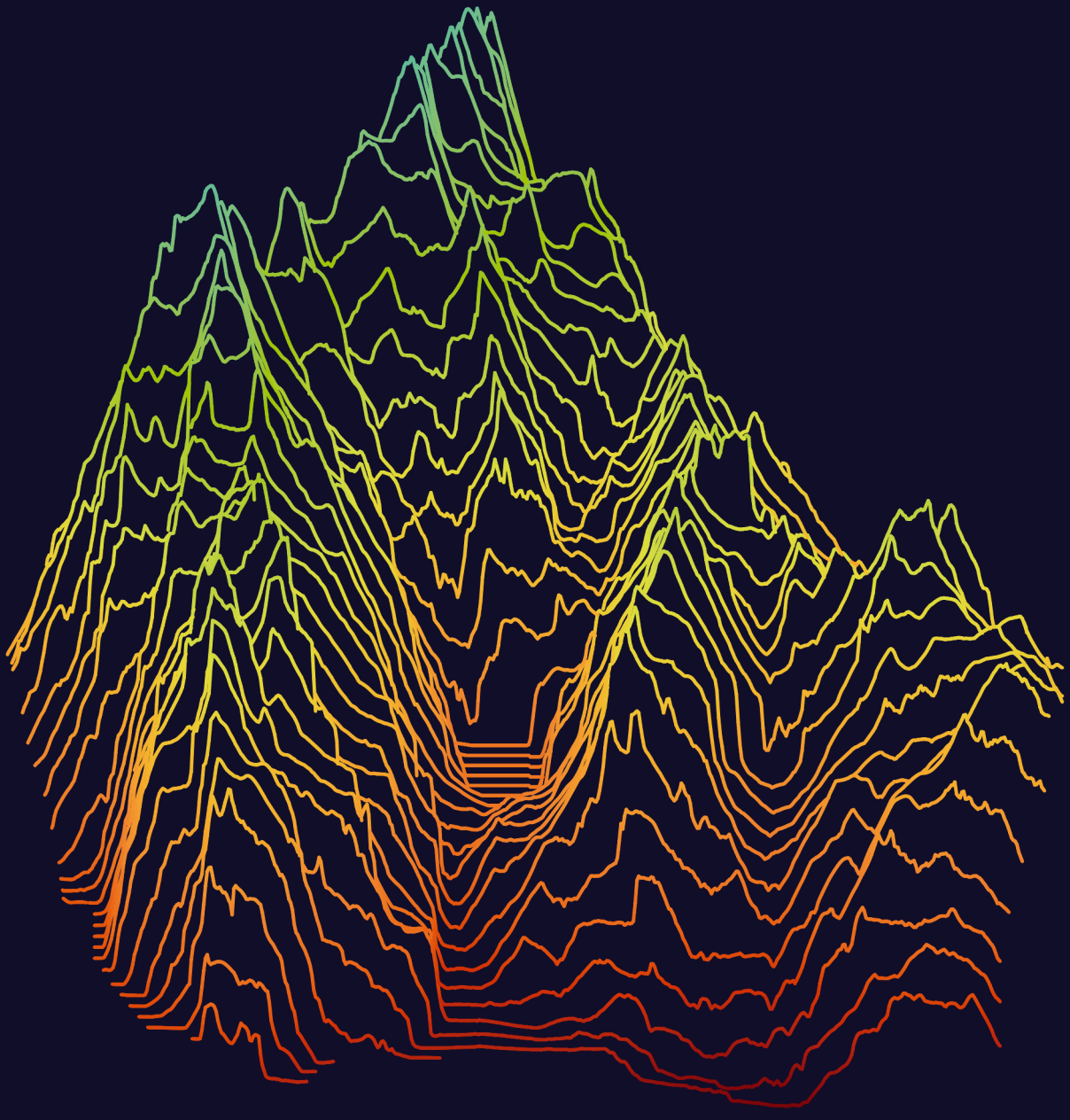
Analysis of sequential treatments for hematological diseases by advanced statistical methods

Koster, E.A.S.



7

Summary and general discussion

SUMMARY

The aim of this thesis was to investigate how analysis of specific clinical settings and application of advanced statistical methodology on high-quality observational data can be used to investigate complex mechanisms and research questions in the field of hematology.

In **Chapter 2**, we aimed to disentangle the effects of competitive repopulation and allo-immunological pressure on the patient- and donor-derived lymphohematopoietic recovery after allogeneic hematopoietic stem cell transplantation (alloSCT). For this, we selected a cohort of 281 patients with acute leukemia receiving alemtuzumab-based T-cell depleted (TCD) alloSCT after myeloablative (MA) or nonmyeloablative (NMA) conditioning. Part of this cohort received a prophylactic donor lymphocyte infusion (DLI) at 3 months after alloSCT because of an anticipated high relapse risk, while the rest of the cohort could receive DLI from 6 months after alloSCT. This setting provided us a natural control and intervention group for the 3-month DLI. We first investigated the recovery before any DLI. Without DLI, the allo-immunological pressure was low: the 3-month cumulative risk of clinically relevant Graft-versus-Host-Disease (GVHD), i.e., GvHD requiring therapeutic systemic immunosuppression (tIS), was 13% (95% confidence interval [95%-CI] 9-17) in the total cohort and only 2% (95%-CI 0-5) after NMA conditioning. Despite the low allo-immunological pressure, 99% of the patients engrafted, showing that primary engraftment did not depend on MA conditioning or the presence of evident allo-immunological pressure. However, the establishment of complete donor-derived hematopoiesis depended on both the conditioning intensity and the presence of allo-immunological pressure: at 3 months, 32% of the NMA-conditioned patients without any GvHD showed full-donor bone marrow (BM) chimerism compared to 71% of the MA-conditioned patients without any GvHD and 88% of the MA-conditioned patients with GvHD. Granulocytes, monocytes, natural killer (NK) cells and B cells closely followed the BM repopulation status. In contrast, even in patients with complete donor-derived hematopoiesis, circulating CD4+ and CD8+ T cells could be predominantly of patient origin. The 3-month level of donor chimerism in these cell populations depended on the conditioning intensity: of the NMA-conditioned patients 7% and 12% had full-donor chimerism (FDC) in CD4+ and CD8+ T cells, respectively, compared to 33% and 41% of the MA-conditioned patients. We did not observe a significant difference between MA-conditioned patients with and without GvHD, which may be explained by the systemic immunosuppressive treatment that the patients with GvHD still received at the time of this measurement. To assess the impact of the introduction of allo-immunological pressure by DLI, we first compared the 3- and 6-month measurements between patients with a 3-month DLI who did not develop GvHD and patients who did not receive a 3-month DLI and did not develop GvHD. The latter group showed stable BM chimerism kinetics in this period (66% FDC at 3 months compared to 61% at 6 months). In contrast, patients of the DLI group often showed conversion to FDC: 38% had FDC at 3 months compared to 63% at 6 months, demonstrating that the 3-month DLI could induce chimerism conversion even in the absence of GvHD. CD4+ and CD8+ T-cell chimerism kinetics showed similar patterns with increasing levels of donor chimerism in the patients with DLI but not in those without DLI, suggesting that for the establishment of a completely donor-derived T-cell compartment, some allo-immunological pressure is needed. Finally, we investigated the

allo-immunological effects of the total DLI strategy in all patients with mixed hematopoiesis at the time of their first DLI. Of the 65 patients, 72% converted to FDC, of whom only 34% developed clinically relevant GvHD. These results illustrate that the Graft-versus-Leukemia (GvL) effect can be separated from GvHD.

In **Chapter 3**, we investigated the complex associations between immune cell kinetics and alloreactivity by joint modeling. We selected the same clinical setting as in **Chapter 2**, except that we only included 166 NMA-conditioned patients in order to have a single conditioning intensity without any post-alloSCT GvHD prophylaxis that might have influenced the immune cell kinetics. First, we investigated the effect of the 3-month DLI on the kinetics of T-cell and NK-cell counts after TCD alloSCT. For this, we constructed a joint model that considered the first 6 months after alloSCT and compared two groups in an intention-to-treat approach: those scheduled for a 3-month DLI because of an anticipated high risk of relapse (the 'high risk' group) and those who were not (the 'non-high risk' group). The model was run separately for the counts of total (CD3+) T cells, CD4+ T cells, CD8+ T cells and NK cells. The clinical events of interest were start of tIS for GvHD, relapse and other failure (i.e., death, graft failure, start of systemic immunosuppression for a non-GvHD indication and virus-specific T-cell infusion for a severe viral infection). Aside from disease risk group the model also considered donor type (related donor [RD] versus unrelated donor [UD] with anti-thymocyte globulin (ATG) additionally to the alemtuzumab) and patient/donor CMV status (both seronegative or not). Compared to patients with a RD, patients with an UD receiving an ATG-containing conditioning regimen had lower T-cell counts during the first 3 months after alloSCT, illustrating the enduring effect of ATG. However, for those with an UD, starting from 3 months the T-cell trajectories started to diverge between the high and non-high risk groups, resulting in higher T-cell counts in those intended to receive a 3-month DLI. As the only plausible explanation for this increase is the 3-month DLI, these data show that DLI can lead to detectable T-cell expansion. Notably, we did not see a divergence between the risk groups with a RD. We observed significantly more GvHD in the high risk group (hazard ratios [HRs] ranging between 6.3 [CD8 model] and 7.3 [CD4 model]). Also higher CD3 and CD4 counts were associated with a higher risk of GvHD (HR per unit log count increase: 2.4 [95%-CI 1.4-4.1] and HR 1.5 [95%-CI 1.0-2.3], respectively). Higher CD4 counts decreased the risk of relapse (HR 0.6, 95%-CI 0.5-0.9) and other failure (HR 0.7, 95%-CI 0.6-1.0). NK cell counts were associated with a higher risk of GvHD and a lower risk of relapse. However, when including both CD4 and NK cell counts in an exploratory time-dependent cause-specific Cox model for GvHD, the effect of NK cell counts disappeared, suggesting that the observed association between NK cell counts and GvHD merely reflected the high correlation between the counts of NK cells and CD4+ T cells and not a direct effect of NK cell counts on the risk of GvHD. To further investigate the T-cell kinetics after the 3-month DLI, we constructed a second joint model starting from this DLI, only including those who actually received this DLI. Having an UD (HRs ranging between 7.0 [CD8 model] and 22.5 [CD4 model]) and higher CD3, CD4 and CD8 counts (HRs ranging between 1.6 [CD8 model] and 6.7 [CD4 model]) were all associated with a higher risk of GvHD during the first 3 months after DLI.

In **Chapter 4**, we aimed to identify other risk factors that influence the alloreactivity of DLI, considering a cohort of patients with acute leukemia receiving their first DLI at 3

(n = 88) or 6 (n = 76) months after alemtuzumab-based TCD alloSCT. First, we assessed the relationship between the timing and dose of DLI and the risk of clinically relevant GvHD in relation to donor and conditioning type. The tenfold dose difference between the 3- and 6-month DLI resulted in similar risks of GvHD: 28% (95%-CI 20-40) and 30% (95%-CI 22-43) at 3 months after the 3- and 6-month DLI, respectively. For both DLIs, the 50% dose reduction in case of an UD sufficed for equalizing the GvHD risks between patients with RD and UD after MA conditioning. In contrast, NMA-conditioned patients with an UD still had a higher risk of GvHD than NMA-conditioned patients with a RD. Then, we focused on three conditions at the time of DLI that could promote T-cell activation: the presence of patient-derived antigen-presenting cells (APCs) as estimated by the BM chimerism status, lymphopenia and the presence of a viral infection close to DLI. As we wanted to estimate the effects of these risk factors on the development of GvHD, the risk of death during GvHD and on the total clinical outcome after DLI, we constructed a time-inhomogeneous Markov multi-state model starting at the time of first DLI and considering the following events: start of tIS for GvHD, stop of tIS, relapse, death and second DLI. The model was run three times for the 3- and 6-month DLI separately, each time including only one of the factors of interest and donor/conditioning type. For the 3-month DLI, viral infections close to DLI increased the risk of GvHD (HR 3.7, 95%-CI 1.7-7.9), while we observed no significant associations with BM chimerism or lymphopenia. At the time of the 6-month DLI, viral infections were uncommon and played no important role in the development of GvHD. Instead, the presence of ≥5% mixed chimerism (MC) in the BM significantly increased the risk of GvHD (HR 3.6, 95%-CI 1.2-11.3) while the presence of 1-5% MC in the BM and lymphopenia showed a trend of increasing the risk of GvHD. We did not observe significant associations between the risk of death during tIS and the main risk factors of GvHD, i.e., viral infections for the 3-month DLI and BM chimerism for the 6-month DLI. To demonstrate the impact of viral infections on current GvHD-relapse-free survival (cGRFS) after the 3-month DLI, we extended the multi-state model and compared the 6-month cGRFS from different starting states: 61% (95%-CI 50-73) from the state 'DLI without viral infection' versus 31% (95%-CI 19-52) from the state 'DLI with viral infection'. For the 6-month DLI, we integrated the two transition-specific Cox models with BM chimerism as components in the multi-state model to predict the outcome for two reference patients, a MA-conditioned patient with a RD and FDC at time of DLI and a MA-conditioned patient with a RD and ≥5% MC. The 6-month cGRFS for these reference patients was 77% (95%-CI 60-98) and 44% (95%-CI 19-100), respectively. The strong impact of viral infections and BM chimerism on cGRFS underline the clinical relevance of these findings.

In **Chapter 5**, we investigated how the transplantation strategy affects the alloreactivity of DLI by considering a different clinical setting, DLI following alloSCT with posttransplant cyclophosphamide (PTCY). Like TCD, PTCY can be applied as partial *in vivo* T-cell depletion early post-transplant to reduce the risk of severe GvHD after alloSCT, but it leads to faster immunological recovery and more FDC early after alloSCT compared to TCD. In this setting, the low-dose DLI was given at 4 months after alloSCT instead of the 3 months in the case of alemtuzumab-based TCD alloSCT. First, we examined the risk factors we had identified in **Chapter 4**. All risk factors were uncommon: of the 83 patients receiving a 4- or 6-month DLI, only 5% had a viral infection close to DLI, 6% had ≥5% mixed BM chimerism and 17% had lymphopenia,

far less that what we had observed in the alemtuzumab setting (19%, 27% and 47%, respectively). We then investigated the development of clinically relevant GvHD after DLI. In line with the low presence of these risk factors, the risk of GvHD was very low: 4% (95%-CI 0-8) at 3 months after DLI. Only one patient died of GvHD, after receiving a 6-month DLI while having 14% patient material in the BM chimerism sample. The combined results of **Chapter 4** and **Chapter 5** indicate that transplantation strategies have a profound impact on the conditions at time of DLI, which in turn influence DLI alloreactivity. To investigate whether DLI after PTCY alloSCT could still induce chimerism conversion from MC to FDC, we examined the BM chimerism kinetics of the 28 patients with MC at time of their first DLI: 79% converted to FDC, of whom only 9% developed clinically relevant GvHD. This conversion rate was similar to what we had observed in **Chapter 2**, while the risk of GvHD was lower. None of the responders relapsed, indicating achievement of a meaningful GvL effect.

In **Chapter 6**, we aimed to investigate whether and how the multi-state framework could be used to develop a comprehensive measure of "treatment success" that can capture the complex clinical recovery and failure patterns of patients with aplastic anemia (AA) receiving immunosuppressive therapy (IST). We defined three levels of treatment success. The broad aim of IST for AA is to achieve and maintain transfusion independency without the development of secondary BM diseases like acute myeloid leukemia (AML) or myelodysplastic syndrome (MDS). We captured this by the endpoint Disease-free survival (DFS). Of note, this endpoint does not consider blood cell counts: it is possible for a patient to be in a DFS state while still having granulopenia. However, transfusion independency is likely more indicative for quality of life than exact blood cell counts. DFS should preferably be achieved without requiring an alloSCT: Transplantation- and Disease-free survival, T-DFS. The ultimate aim is to stop all AA therapy after a response is achieved, which is captured by the endpoint Transplantation-, Treatment- and Disease-free survival, TT-DFS. We estimated these endpoints in a cohort of 127 transfusion-dependent patients with AA using a time-inhomogeneous Markov multi-state model to allow for gain, loss and recovery of response. The 5-year probabilities of DFS, T-DFS and TT-DFS were 70% (95%-CI 61-81), 60% (95%-CI 51-71) and 42% (95%-CI 33-54), respectively. In the next step, we investigated the effects of age, AA severity and the presence of a GPI-deficient cell clone on different transitions. As GPI-deficient cells are considered to be less sensitive to an immune attack and can emerge as a detectable cell population after the other HSCs have been depleted by autoimmune cells, the presence of such a clone can indicate that the AA was caused by autoimmunity. Since only those with AA caused by an autoimmune response are likely to respond to IST, the presence of a GPI-deficient cell clone may be associated with achieving a response. Indeed, having a GPI-deficient cell clone of ≥1% increased the hazard of becoming transfusion-independent (HR 2.2, 95%-CI 1.4-3.4), while age of 40 or above and having severe or very severe AA decreased the hazard with HRs ranging between 0.4 and 0.5. We did not observe significant effects of these risk factors on the likelihood of being able to stop all non-transplant therapy after having become transfusion-independent. As expected, age of 60 or above was a strong predictor for the risk of death (HR 7.3, 95%-CI 1.5-34.3). To demonstrate the impact of these risk factors on the outcomes, we calculated model-based prognoses for reference patients with different baseline characteristics. For example, the model-based 5-year probability of TT-DFS for a patient of 40 years or younger with severe aplastic anemia and no GPI-

deficient cell clone was 47% (95%-CI 33-68), compared to 24% (95%-CI 14-41) for a patient of at least 60 years old with the same characteristics, and 41% (95%-CI 26-65) for a patient of at least 60 years old with a GPI-deficient cell clone and non-severe AA. These results indicate that the three risk factors have a strong impact on the probability of long-term treatment success.

GENERAL DISCUSSION

In this thesis, we demonstrated how detailed observational data and advanced statistical methods can be used to answer clinical and immunological research questions in the field of hematology by capturing complex recovery and failure patterns and their underlying mechanisms. The use of cohorts of patients with or without DLI after different transplantation strategies allowed us to investigate the impact of the transplantation strategy and DLI on the lymphohematopoietic and clinical recovery and to disentangle the roles of competitive repopulation and allo-immunological pressure. Via joint modelling we could quantify the DLI effect on immune cell counts and the associations between the immune cell counts and different clinical events. With the Markov multi-state framework we could investigate the effects of risk factors on different components of the recovery and failure process and translate these to clinically relevant outcome measures. Taking advantage of the versatility of the multi-state framework, we constructed a single comprehensive multi-state model that could estimate the probabilities of three different levels of treatment success after IST for AA over time.

Vital importance of setting of DLI

In **Chapters 4 and 5**, we aimed to investigate the effects of factors that might increase the risk of DLI-induced GvHD: the presence of patient-derived APCs, a proinflammatory environment and cytokines that can amplify GvHD. However, none of these factors were measured standardly. Instead, for each of the actual factors of interest we considered a measure that was available as a proxy: the presence of MC in the BM, viral infection (one of the most common causes of inflammation) and lymphopenia (during which the concentrations of cytokines that promote T-cell proliferation are higher), respectively. Because these had been measured as part of the standard clinical care, we could use the full cohorts for our analyses. We demonstrated that the presence of mixed chimerism and viral infections can have a strong impact on the risk of DLI-induced GvHD, and that their presence depends on the transplantation strategy and timing of DLI.

Current DLI recommendations do not take conditions at the time of DLI into account aside from strongly advising against the use of DLI to boost the GvL effect in the presence of GvHD and uncontrolled infections.[1,2] This can partly explain the wide range of reported risks of GvHD[3]. More personalized DLI protocols that consider more conditions at the time of DLI would likely improve the balance between efficacy (GvL effect) and toxicity (GvHD) of prophylactic and preemptive DLI. Firstly, the recommendations for DLI in case of infection could be extended. Postponing a DLI until an infection has been cleared is not always feasible, for instance when MRD is increasing rapidly. As an alternative, the DLI dose could be reduced. If an infection occurs just after

DLI, one could more aggressively start tIS upon signs of GvHD. However, the benefit of reducing the risk of severe GvHD should be weighed against the disadvantage of suppressing the immune system while fighting an infection. Secondly, the presence of lymphopenia or MC in the BM could be considered for the dose determination in the preemptive and prophylactic setting. Reducing the dose in case of MC may feel counterintuitive, since patients with MC are more likely to develop a relapse. However, as long as there are no signs of relapsing disease, the presence of MC merely reflects the absence of a GvL effect; the magnitude of MC should not have an effect on the relapse risk. In contrast, a patient with 5% MC likely has more patient-derived APCs and thereby a larger GvHD risk than a patient with 1% MC. Following this reasoning, one would recommend a lower dose for the former patient to minimize the risk of severe GvHD. If no response occurs, further DLIs can be given following a dose-escalating approach. Conversely, for patients with FDC and therefore likely having fewer patient-derived professional APCs, one could increase the DLI dose if they are considered to have a high risk of relapse. None of the FDC patients in our study died of GvHD, suggesting that there is room for increasing the dose. Thirdly, for patients who need a stronger alloimmune response one could *initiate* a proinflammatory condition by administering lymphodepleting chemotherapy before DLI. Miller et al. showed that this is an effective method to increase the alloreactivity of DLI, but had to stop the trial because it caused too much GvHD-related toxicity.[4] Guillaume et al. applied lower doses of lymphodepleting chemotherapy and observed no toxicity in patients receiving DLI after 3 days of 25mg/m^2 fludarabine.[5] While these studies took place in the setting of therapeutic DLI with small patient cohorts, they suggest that low-dose lymphodepletion can be an effective tool to increase DLI efficacy.

Before any of these suggestions can be implemented, they need to be tested and validated in other cohorts and settings. Two types of studies are needed: observational studies to validate the estimated effects of risk factors for GvHD after DLI and intervention studies to investigate dose adjustments based on the conditions at the time of DLI and the use of lymphodepletion in situations where a prompt alloimmune response is needed.

Added value of complex statistical analyses

Two advanced statistical methods were applied in this thesis: Markov multi-state modeling for investigating complex sequences of events and joint modelling for analyzing the trajectories of biomarkers and their effects on survival outcomes. The multi-state models constructed in this thesis showcase several advantages of multi-state modelling. Firstly, because multi-state models can consider sequences of events and analyze events in continuous time without assuming a constant hazard, they can capture dynamic measures of treatment success such as cGRFS and TT-DFS. cGRFS differs from GvHD-relapse-free survival (GRFS) by considering recovery after GvHD: patients in whom GvHD does not resolve or who die of GvHD remain in a failure state, while those who recover move on to a non-failure state. This better reflects the clinical situation during follow-up, since patients with resolved GvHD can have comparable quality of life compared to those who never developed GvHD. [6] Moreover, patients with resolved GvHD may benefit from the concomitantly established GvL effect reducing their risk of relapse, as shown in **Chapter 4** (none of the patients who started tIS for GvHD relapsed) and by others[7]. In the setting of IST for AA, we defined three dynamic

endpoints, the DFS, T-DFS and TT-DFS, to evaluate different levels of treatment success (**Chapter 6**). In contrast to the commonly used 6-month response rate and overall survival, these measures shows the loss and gain of different levels of response over time and give information on the different treatments of AA during follow-up and the risks of death and development of secondary BM diseases. Secondly, the approach of including transition-specific risk factors focuses on the underlying biological processes. This is more logical than estimating the effect of risk factors directly on composite endpoints such as GRFS. For instance, Tan et al. [8] constructed an extensive multivariable Cox model for GRFS. While such an approach may be sufficient if one is only interested in the prediction of GRFS, it does not help to understand why certain risk factors are important: donor type was not significant, but does this mean that it was irrelevant or that the opposing effects of higher genetic disparity on relapse and GvHD cancelled each other out? Several studies have shown the added value of multi-state modelling by reanalyzing trials.[9-11] For instance, Bakunina et al. reanalyzed a trial[12] where patients with AML were randomized to receive remission induction therapy with or without clofarabine, showing that clofarabine reduced the risk of relapse but did not improve survival. Their multi-state approach enabled them to consider intermediate events such as consolidation by alloSCT and achievement of MRD negativity, and showed that the addition of clofarabine reduced the risk of relapse irrespective of MRD status or alloSCT, but increased the risk of non-relapse mortality before alloSCT.[10] In **Chapter 6**, the transition-specific age effects in the multi-state model on patients with AA receiving IST showed that the relatively poor TT-DFS of patients aged 60 years or older can be explained by both a lower hazard of achieving a response and a higher hazard of death, of which a descriptive analysis indicated that the majority was not related to treatment toxicity but to pancytopenia. The possibility to use the estimated transition-specific effects to calculate model-based prognoses and thereby show the total impact on the clinical outcome is the third advantage of multi-state modelling. Of note, predictions can be given for each state separately or for combinations of states, allowing to show the impact on several clinical outcomes of interest (**Chapters 4 and 6**). For treatment decisions and prognosis, outcome predictions are often more relevant than HRs, as they can take the full recovery process and opposing effects of the same risk factor on different transitions into account as well as the baseline hazards.

There are also some limitations to the multi-state models used in this thesis. They depend on the Markovian assumption, meaning that the risk to make a certain transition only depends on the state, the time since start of the analysis and, if estimated, the transition-specific covariate effects. This is a simplification, as for instance the risk of death after relapse likely also depends on the timing of the relapse (early relapses usually have a worse prognosis than later relapses) and the time since relapse (e.g., those who are still alive a year after relapse likely have a lower mortality risk than those who just developed a relapse). Multiple timescales have only been implemented in parametric models[13], which require more assumptions than non- and semi-parametric models. As an alternative, states can be split into multiple states to include some of the information of the time since the start of a clinical event. For instance, continuing with the relapse example, relapse could be split into early and late relapse and additional states could be added such as '1 year after relapse'. However, this approach can quickly lead to very extensive multi-state models requiring more data in order to have sufficient events for the transitions. Moreover, interpreting covariate effects would be more complex, as separate

coefficients are calculated for each transition. Because of these reasons, we did not apply this approach in the studies in this thesis, even though for example in **Chapter 6**, the likelihood of stopping all non-transplant therapy after having become transfusion-independent likely depends on both the time since start (as according to the Dutch AA guidelines IST should be given for at least 6 months) and on the time since achieving transfusion independency.

The other advanced model applied in this thesis is the joint model. In **Chapter 3**, the raw data (see Supplemental Figure 3) did not hint at a difference in the T-cell count trajectories between those who eventually did or did not develop GvHD, but the association could be made visible and quantified using joint modelling. This method is far more efficient than landmarking such as performed by Podgorny et al.[14], as landmarking required them to exclude patients who had developed GvHD before the landmark time and to consider the immune cell counts at the landmark time as fixed baseline covariates for their Cox proportional hazards model. As outlined in **Chapter 1**, joint models can model the trajectories of biomarkers over time without assuming constant values between measurements or absence of measurement error. Moreover, joint models can yield individualized predictions to visualize and quantify how changes in the biomarker values affect the risk of clinical events, as illustrated by Baart et al.[15] As an example they predicted the risk of neo-aortic valve regurgitation for two patients who underwent surgery shortly after birth because of transposition of the great arteries, updating their risk each time the neo-aortic root diameter was measured (i.e., dynamic prediction). While at the start both patients had the same risk of this event, their risks diverged considerably over time as in one patient the root diameter increased more slowly than in the other patient. Dynamic prediction tools like this can have a great value in the clinical follow-up of patients.

Barriers for widespread application of complex statistical methods

During the studies explored in this thesis, we encountered several limitations due to the relatively low numbers of patients and events compared to the complexity of our models. The joint model in **Chapter 3** was based on 166 patients and considered four immune cell populations and three endpoints of interest, GvHD, relapse and a composite of all other failures. As GvHD and relapse depend on the presence and absence of allo-immunological pressure, respectively, we expected opposite effects of the immune cell counts on these events and included them as separate endpoints. The third endpoint was needed to stop the follow-up as soon as an event occurred that could influence the immune cell counts or the risks of relapse and GvHD (aside from the 3-month DLI, which was the intervention of interest). As incorporation of all four immune cell populations in a single model would require far more assumptions regarding the association structure, we had to investigate the immune cell populations of interest in separate models, which was a lot considering the size of our dataset. In **Chapter 4**, we had 88 patients with a 3-month DLI and 76 with a 6-month DLI, which were analyzed in two separate multi-state models with 14 states. In both models, about 30 patients made the transition from DLI1 to start of tIS for GvHD. The three risk factors of interest had to be analyzed in separate Cox models as we needed to include conditioning/donor combination, which was associated with both the presence of the risk factors and the risk of GvHD after DLI. Having more events would have allowed us to include all our risk

factors of interest in one comprehensive model. The numbers of patients who died during tIS were even lower (16 and 12 for the 3- and 6-month DLI, respectively), which likely explained why we did not detect significant effects of high MC or viral infections on the risk of death during tIS for GvHD. Moreover, none of the patients with FDC died during tIS for GvHD, necessitating us to use the presence of either FDC or low MC as the reference category for high MC. Therefore, we could not quantify the increased hazard of death during tIS for patients with high MC compared to patients with FDC. In **Chapter 6** we had a larger cohort, but the low numbers of events for the transitions to death required us to assume a shared baseline hazard in order to assess the effect of age on the risk of death. Thus, while our analyses yielded valuable results, they were less precise and required more assumptions than if we would have had more events. In order to construct even more complex models or get more precise estimates, larger datasets are needed. However, multi-state and joint models require high-quality detailed observational data, which is often only collected in relatively small cohorts. Registries have large cohorts, but will most likely need to improve their data collection in order to have data of sufficient quality and detail for these types of models. This requires more commitment of the registries and their participating centers.

Another barrier for the use of complex statistical methods such as multi-state and joint models is the required level of statistical knowledge to construct these models and to correctly interpret the results. **Chapter 3** was a joint project of the department of Hematology and the department of Biomedical Data Sciences to ensure sufficient knowledge both on the clinical and immunological processes and on the modelling techniques. Assuming a shared baseline hazard for different transitions as we did in **Chapter 6**, could only be done after carefully considering the medical/biological implications of this statistical assumption. Moreover, having a correct model does not guarantee correct interpretation. An audience with less experience in multi-state methodology may find it difficult to for instance understand why risk factors can have opposing effects on different transitions in a multi-state model or to understand the implications of all model assumptions.

Outlook

Our studies show that the combination of detailed data and advanced statistical methods can be used to answer complex research questions using real-world clinical data. Standard collection of detailed observational data has several advantages over data collected for a specific study. Firstly, standardly collected data that are stored in a single place can be easily used for multiple studies. It also provides a reliable data source for all consecutive patients that can be accessed rapidly in the case of fast-changing developments. For instance, the unexpected change in DLI-induced GvHD risk after switching from alemtuzumab-based TCD to PTCY alloSCT could be investigated early after implementation of PTCY, because in both settings detailed information was collected in a standardized way. Ethically, data may only be collected for research if they are needed to answer relevant research questions. To this end, it is essential that both before the start of and during the data collection, researchers, data managers and methodologists discuss which research questions may be of interest, what measurements could be useful and analyzable and how they can be most efficiently collected. Also for

data being collected for clinical care, it is valuable to discuss with a statistician whether the data are usable for analysis.

If more data with sufficient detail and quality become available, it will be possible to extend the models used in this thesis or apply them in other settings. For the AA model, more covariates could be included, and we could also include covariates on other transitions, such as relapse after having achieved transfusion independency. The observed differences between the alemtuzumab-based TCD and PTCY setting suggest that it would be very valuable to repeat the analyses in different alloSCT settings to validate the results. For all multi-state models, it would also be possible to incorporate biomarkers, for instance T-cell counts in the cGRFS model and blood counts in the AA model. This approach has already been implemented by Ferrer et al.[16], but has not yet been applied in the field of hematology. As mentioned before, large datasets are mainly available from registries. It is unlikely that very large registries such as the European Society for Blood and Marrow Transplantation will collect very detailed clinical data or biomarker data. Collaboration projects between centers and regional or national registries are more likely to obtain sufficient data of the required quality, since a lower number of participating centers usually means higher commitment for precise and detailed data collection.

After the models have been validated, they may be of value in clinical practice. Model predictions can help with treatment decisions, shared decision making and counseling of the patients. A major advantage of both the multi-state model and the joint model is that they can incorporate new information such as biomarker trajectories and the occurrences of clinical events during follow-up. This allows to update the prognosis of patients based on their trajectory so far, which may help in the decision for further treatment lines. The models in this thesis were constructed to better understand underlying mechanisms behind recovery and failure patterns and not to serve as a prediction tool. For that, the results of the semi-parametric analyses need to be validated in large cohorts, that are preferably treated according to the current treatment guidelines. For instance, the recent addition of eltrombopag to the first-line IST treatment protocol for patients with AA may affect both the prognosis and the effect sizes of the risk factors: the estimated values in **Chapter 6** might not hold for patients that are treated according to the new guideline.

Finally, to narrow the gap between clinical researchers and methodologists, papers applying advanced statistical methods should become more accessible to a less methodological audience. To serve both audiences, it is important to explain the methods both on a more general level and a more technical level. In the studies of this thesis that report multi-state models, we discussed the model results in steps: first the non-parametric analyses to show how the transition probabilities changed over time, then the Cox models per transition and finally predictions for reference patients. By analyzing clinically relevant questions with real-world data, publishing in clinical or immunological journals and explaining the methods comprehensively, not only methodologists but also clinical researchers can be reached. Hopefully, by showing by example the added value of more advanced statistical models in answering important questions, the collaboration and cross-talk between the different fields will increase, leading to more opportunities for methodologists to develop or apply methods on clinically relevant applications and for clinical researchers to answer complex research questions.

REFERENCES

1. Falkenburg, J. H. F., Schmid, C., Kolb, H. J. & Kuball, J. in The EBMT Handbook: Hematopoietic Cell Transplantation and Cellular Therapies (eds Anna Sureda et al.) 531-539 (Springer International Publishing, 2024).
2. Pagliuca, S., Schmid, C., Santoro, N. et al. Donor lymphocyte infusion after allogeneic haematopoietic cell transplantation for haematological malignancies: basic considerations and best practice recommendations from the EBMT. The Lancet Haematology 11, e448-e458 (2024). doi: 10.1016/S2352-3026(24)00098-X
3. Biederstädt, A. & Rezvani, K. How I treat high-risk acute myeloid leukemia using preemptive adoptive cellular immunotherapy. Blood 141, 22-38 (2023). doi: 10.1182/blood.2021012411
4. Miller, J. S., Weisdorf, D. J., Burns, L. J. et al. Lymphodepletion followed by donor lymphocyte infusion (DLI) causes significantly more acute graft-versus-host disease than DLI alone. Blood 110, 2761-2763 (2007). doi: 10.1182/blood-2007-05-090340
5. Guillaume, T., Gaugler, B., Chevallier, P. et al. Escalated lymphodepletion followed by donor lymphocyte infusion can induce a graft-versus-host response without overwhelming toxicity. Bone Marrow Transplantation 47, 1112-1117 (2012). doi: 10.1038/bmt.2011.231
6. Kurosawa, S., Yamaguchi, T., Oshima, K. et al. Resolved versus Active Chronic Graft-versus-Host Disease: Impact on Post-Transplantation Quality of Life. Biol Blood Marrow Transplant 25, 1851-1858 (2019). doi: 10.1016/j.bbmt.2019.05.016
7. Stern, M., de Wreede, L. C., Brand, R. et al. Sensitivity of hematological malignancies to graft-versus-host effects: an EBMT megafile analysis. Leukemia 28, 2235-2240 (2014). doi: 10.1038/leu.2014.145
8. Tan, J., Wang, Y., Yu, S. J. et al. Prognostic factors on graft-versus-host disease-free and relapse-free survival after allogeneic hematopoietic stem cell transplantation for adults with acute leukemia. Leuk Res 59, 1-7 (2017). doi: 10.1016/j.leukres.2017.05.011
9. Putter, H., van der Hage, J., de Bock, G. H., Elgalta, R. & van de Velde, C. J. Estimation and prediction in a multi-state model for breast cancer. Biometrical journal. Biometrische Zeitschrift 48, 366-380 (2006). doi: 10.1002/bimj.200510218
10. Bakunina, K., Putter, H., Versluis, J. et al. The added value of multi-state modelling in a randomized controlled trial: The HOVON 102 study re-analyzed. Cancer medicine 11, 630-640 (2022). doi: 10.1002/cam4.4392
11. Hermans, S. J. F., van Norden, Y., Versluis, J. et al. Benefits and risks of clofarabine in adult acute lymphoblastic leukemia investigated in depth by multi-state modeling. Cancer medicine 13, e6756 (2024). doi: 10.1002/cam4.6756
12. Löwenberg, B., Pabst, T., Maertens, J. et al. Therapeutic value of clofarabine in younger and middle-aged (18-65 years) adults with newly diagnosed AML. Blood 129, 1636-1645 (2017). doi: 10.1182/blood-2016-10-740613
13. Iacobelli, S. & Carstensen, B. Multiple time scales in multi-state models. Statistics in Medicine 32, 5315-5327 (2013). doi: 10.1002/sim.5976
14. Podgorny, P. J., Liu, Y., Dharmani-Khan, P. et al. Immune cell subset counts associated with graft-versus-host disease. Biol Blood Marrow Transplant 20, 450-462 (2014). doi: 10.1016/j.bbmt.2014.01.002
15. Baart, S. J., van der Palen, R. L. F., Putter, H. et al. Joint Modeling of Longitudinal Markers and Time-to-Event Outcomes: An Application and Tutorial in Patients After Surgical Repair of Transposition of the Great Arteries. Circ Cardiovasc Qual Outcomes 14, e007593 (2021). doi: 10.1161/circoutcomes.120.007593
16. Ferrer, L., Rondeau, V., Dignam, J. et al. Joint modelling of longitudinal and multi-state processes: application to clinical progressions in prostate cancer. Stat Med 35, 3933-3948 (2016). doi: 10.1002/sim.6972

7