Computational methods to detect HLA-associated mutations

Inaugural-Dissertation
zur
Erlangung des Doktorgrades
Dr. rer. nat.

der Fakultät für
Biologie
an der

Universität Duisburg-Essen

vorgelegt von

Daniel Habermann

aus Wuppertal

April 2022

Die der vorliegenden Arbeit zugrunde liegenden Experimente wurden in den Abteilungen für Bioinformatik des Zentrums für Medizinische Biotechnologie (ZMB) der Universität Duisburg-Essen durchgeführt.

1. Gutachter: Prof. Dr. Daniel Hoffmann

2. Gutachterin: Prof. Dr. Katharina Fleischhauer

Vorsitzende des Prüfungsausschusses: Prof. Dr. Elsa Sánchez-García
Tag der mündlichen Prüfung: 29.08.2022

DOI: 10.17185/duepublico/77062
URN: urn:nbn:de:hbz:465-20230404-083737-3

Contents

Summary

Viruses replicate intracellularly, which means that they are well hidden from the humoral arm of the immune system. However, evolution brought up mechanisms to combat viral infections: All cells with active protein biosynthesis continuously present peptides on the cell surface via Human Leukocyte Antigen (HLA) molecules. Cytotoxic T cells are selected to not bind to peptides originating from the human proteome, but –facilitated by the vast T cell receptor repertoire– may bind to viral epitopes and induce killing of the virus infected cell. In this way, cytotoxic T cells are sharp weapons to combat viral infections and exert strong selection pressure towards virus variants that escape that immune recognition pathway.

HLA escape may occur through many mechanisms, for example through point mutations that reduce binding of the epitope to HLA molecules. The identification of these HLA-associated mutations (HAMs) is not only important for understanding viral evolution, but also impacts the development of broadly effective anti-viral treatments and vaccines against variable viruses. Unfortunately, experimental methods to detect HAMs are prohibitively expensive and too time-consuming for large-scale use. A promising alternative are methods that detect HAMs through the statistical analysis of viral sequence data annotated with host HLA information. Existing methods fail to take confounding effects like phylogeny and important prior knowledge like epitope prediction into account. This necessitates an improved model, which accounts for confounding effects and combines as much information as possible into a single coherent statistical model.

I introduce this thesis with a brief description of the immunological concepts that are important to understand the challenges of identifying HLA-associated mutations through statistical analysis (chapter 1) and an introduction to Bayesian modeling (chapter 2), which provides the statistical basis for the main work. Chapter 3 provides the main results of this thesis, which are summarized in two publications. Chapter 4 concludes with a discussion of the contributed articles.

In the first publication (section 3.1), I present `HAMdetector`, a regression model to identify HLA-associated mutations in HLA-annotated viral sequence data. The main feature of this model is the Bayesian framework, which allows including prior information in a principled way and takes sources of uncertainty into account. The model includes knowledge about the distributional properties of HLA-associated mutations and the fact that HAMs preferentially lie within the boundary of epitopes, which can be predicted

using epitope prediction software. On a large collection of HIV, HBV and HDV datasets, `HAMdetector` identified many potential HAMs that are currently unknown, which implies that a significant fraction of interactions between viruses and T cell based immunity is yet to be discovered, an exciting discovery for immunology and virology.

In the second publication, I transfer the general statistical principle of including as much information as possible in a single coherent model to the field of cancer research: Merkel cell carcinoma is an aggressive type of skin cancer, which can be treated using novel antibody-based therapies. However, these therapies are sometimes not effective, and the risk factors leading to therapeutic failure are not well understood. Using a relatively small dataset of 114 patients characterized by therapy outcome on an ordinal scale (progressive disease, stagnant disease, partial response and complete response), I could show in this collaborative project that out of 17 different patient- and tumor characteristics, immunosuppression and spread of the tumor to multiple organs appear to be linked most strongly to treatment non-response. The main feature of this model is that it takes the ordinal nature of the response into account, an important piece of information that is often discarded in statistical models. This work puts a strong emphasis on model testing, highlighting benefits of the Bayesian workflow to learn as much from the available data as possible and accurately account for model uncertainty.

Zusammenfassung

Viren vermehren sich intrazellulär und sind vor dem humoralen Arm des Immunsystems deshalb gut verborgen. Die Evolution hat jedoch Mechanismen zur Bekämpfung von Virusinfektionen hervorgebracht: Alle Zellen mit aktiver Proteinbiosynthese präsentieren über Human Leukocyte Antigen (HLA) Moleküle ständig Peptide auf der Zelloberfläche. Zytotoxische T-Zellen sind so selektiert, dass sie nicht an Peptide aus dem menschlichen Proteom binden, sondern –unterstützt durch das große Repertoire an T-Zell-Rezeptoren– an virale Epitope binden können und die Abtötung der virusinfizierten Zelle induzieren. Auf diese Weise sind zytotoxische T-Zellen scharfe Waffen zur Bekämpfung von Virusinfektionen und üben einen starken Selektionsdruck auf Virusvarianten aus, die diesem Immunmechanismus entgehen.

Ein solcher HLA-Escape kann durch viele verschiedene Mechanismen erfolgen, zum Beispiel durch Punktmutationen, die die Bindung des Epitops an HLA-Moleküle verringern. Die Identifizierung dieser HLA-assoziierten Mutationen (HAMs) ist nicht nur wichtig für das Verständnis viraler Evolution, sondern hat auch Auswirkungen auf die Entwicklung von wirksamen antiviralen Behandlungen und Impfstoffen gegen variable Viren. Leider sind experimentelle Methoden zum Nachweis von HAMs unerschwinglich teuer und zu zeitaufwändig für einen breiten Einsatz. Eine vielversprechende Alternative sind Methoden, die HAMs durch die statistische Analyse von mit Wirts-HLA-Informationen annotierten viralen Sequenzdaten aufspüren. Vorhandene Methoden berücksichtigen Effekte wie Phylogenie und wichtiges Vorwissen wie Epitopvorhersagen nicht. Dies macht ein besseres Modell erforderlich, das Störvariablen berücksichtigt und so viele Informationen wie möglich in einem kohärenten statistischen Modell zusammenfasst.

Ich leite diese Arbeit mit einer kurzen Beschreibung der immunologischen Konzepte ein, die wichtig sind, um die Herausforderungen bei der Identifizierung von HLA-assoziierten Mutationen durch statistische Analysen zu verstehen (Kapitel 1), sowie eine kurze Einführung in die Bayes'sche Modellierung, die die statistische Grundlage für die Hauptarbeit bildet. Kapitel 3 enthält die Hauptergebnisse dieser Arbeit, die in zwei Veröffentlichungen zusammengefasst sind. Kapitel 4 umfasst eine Diskussion dieser Ergebnisse.

In der ersten Veröffentlichung (Abschnitt 3.1) stelle ich `HAMdetector` vor, ein Regressionsmodell zur Identifizierung von HLA-assoziierten Mutationen in HLA-annotierten viralen Sequenzdaten. Das Hauptmerkmal dieses Modells ist das Bayes'sche Framework, das es erlaubt, Vorwissen direkt in das Modell einzubeziehen und Quellen von Unsicherheit

zu berücksichtigen. Das Modell nutzt Informationen über die Verteilungseigenschaften von HLA-assoziierten Mutationen und die Tatsache, dass HAMs bevorzugt innerhalb von Epitopen liegen, welche mit Hilfe von Epitopvorhersagesoftware vorhergesagt werden können. Anhand einer großen Sammlung von HIV-, HBV- und HDV-Datensätzen identifiziert `HAMdetector` viele derzeit noch unbekannte HAMs, was darauf hindeutet, dass ein erheblicher Teil der Wechselwirkungen zwischen Viren und der T-Zell-basierten Immunität noch nicht entdeckt wurde - eine spannende Entdeckung für die Immunologie und Virologie.

In der zweiten Veröffentlichung übertrage ich das allgemeine statistische Prinzip, so viele Informationen wie möglich in ein kohärentes Modell einzubeziehen, auf den Bereich der Krebsforschung: Das Merkelzellkarzinom ist eine aggressive Form von Hautkrebs, die mit neuartigen antikörperbasierten Therapien behandelt werden kann. Diese Therapien sind jedoch nicht immer wirksam, und die Risikofaktoren, die zu einem Therapieversagen führen, sind nicht gut erforscht. Anhand eines relativ kleinen Datensatzes von 114 Patienten, bei denen der Therapieausgang auf einer ordinalen Skala erfasst wird (fortschreitende Erkrankung, stagnierende Erkrankung, teilweises Ansprechen und vollständiges Ansprechen), konnte ich in diesem Kollaborationsprojekt zeigen, dass von 17 verschiedenen Patienten- und Tumormerkmalen die Immunsuppression und die Ausbreitung des Tumors auf mehrere Organe am stärksten mit dem Nichtansprechen auf die Behandlung zusammenhängen. Das Hauptmerkmal dieses Modells besteht darin, dass es die ordinale Skala der Beobachtungen berücksichtigt, eine wichtige Information, die in statistischen Modellen oft unberücksichtigt bleibt. Diese Arbeit legt einen starken Schwerpunkt auf das Testen von Modellen und verdeutlicht die Vorteile des Bayes'schen Workflows, der es erlaubt so viel wie möglich aus den verfügbaren Daten zu lernen und Modellunsicherheit zu berücksichtigen.

Chapter 1

The HLA system

Viral replication requires synthesis of viral proteins. One mechanism how virus-infected cells can be identified is through continuous survey of intracellular proteins: All nucleated cells present peptides via Human Leukocyte Antigen (HLA) molecules on the cell surface, which function as snapshots of currently expressed proteins. In this way, these peptides act as possible antigens to cytotoxic T cells, which are selected to not bind to peptides originating from the usual human proteome, but may bind to epitopes stemming from viral proteins. The pathway from intracellular protein to presented epitope involves antigen processing and presentation and consists of multiple steps (figure 1.1):

1. Proteasomal degradation of intracellular proteins
2. Loading of peptides onto HLA molecules
3. Transport of the HLA/peptide complex to the cell surface

In the next section, these three steps are described in greater detail (section 1.1). The following sections in this chapter describe the molecular properties of HLA molecules and the nomenclature of HLA alleles (section 1.2), the development of cytotoxic T cells (section 1.3), and the mechanisms by which cytotoxic T cells kill virus-infected cells (section 1.4).

1.1 Antigen processing

1.1.1 Proteasomal degradation of intracellular proteins

Cytosolic proteins have a limited half-life, typically ranging from 10 to 140 hours (Cambridge et al., 2011), and are continuously synthesized and degraded. The degradation is achieved by specialized enzyme complexes called proteasomes (Adams, 2003), which cleave proteins at cleavage sites (Saxová et al., 2003) after they have been marked for degradation, usually by appending poly-ubiquitin chains, a cascade-like process initiated by ubiquitin activating enzymes (Buetow and Huang, 2016; Grice and Nathan, 2016).

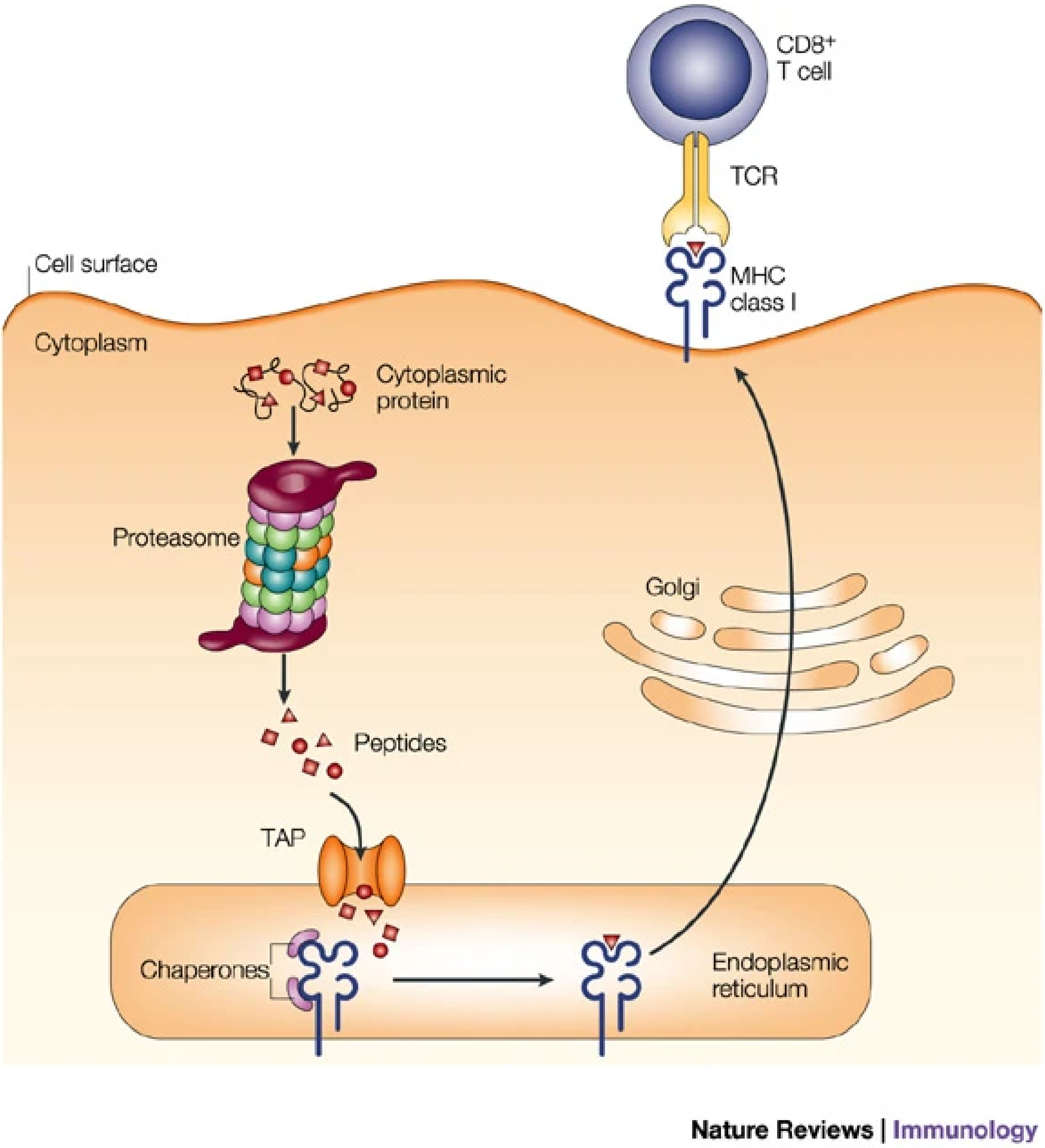


Figure 1.1: Cytosolic and nuclear proteins are degraded by the proteasome into peptides. The transporter for antigen processing (TAP) then translocates peptides into the lumen of the endoplasmic reticulum (ER) while consuming ATP. MHC class I heterodimers wait in the ER for the third subunit, a peptide. Peptide binding is required for correct folding of MHC class I molecules and release from the ER and transport to the plasma membrane, where the peptide is presented to the immune system. TCR, T-cell receptor.

Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Making sense of mass destruction: quantitating MHC class I antigen presentation, Yewdell, J., Reits, E. & Neefjes, J. (2003)

The proteasome complex consists of several subunits, which carry out important functions (Bard et al., 2018) like binding to ubiquitinated proteins (ADRM1, Husnjak et al. (2008)), ATP-mediated unfolding of proteins (PSMC1-PSMC6, Martin et al. (2008)) and cleavage of the substrate (20S core, Zwickl et al. (1999)). The proteasome is shaped like a tunnel (Groll et al., 1997) and the interior of that tunnel provides a finely controlled environment to facilitate cleavage of the substrate (Borissenko and Groll, 2007).

Recent findings suggest that proteasomes also splice peptides generated from the same protein. These proteasome-generated spliced peptides account for about one-fourth of the entire HLA class I immunopeptidome in terms of abundance (Liepe et al., 2016).

After proteasomal degradation, the resulting peptides have a length of about 3 to 25 amino acids (Kisselev et al., 1999). Most of them are further processed to amino acids and used for protein biosynthesis (Vabulas, 2007). However, some peptides are, possibly after further processing (Reits et al., 2004), translocated into the endoplasmatic reticulum and loaded onto HLA molecules.

1.1.2 Loading of peptides onto HLA molecules

The translocation of cytosolic peptides into the endoplasmatic reticulum is facilitated by a specialized hetero-dimer called the transporter associated with antigen processing (TAP, (Abele and Tampé, 2004)). It is ATP driven (Androlewicz et al., 1993; Neefjes et al., 1993; Shepherd et al., 1993) and mostly transports peptides with a length between 8 and 16 amino acids (van Endert et al., 1994), although peptide lengths of 6 to 30 amino acids have also been reported (Koopmann et al., 1996). Peptides with aromatic, hydrophobic or positively charged amino acids near the C-terminus are preferred (Uebel et al., 1997), but the TAP complex lacks any true peptide-specificity (Androlewicz and Cresswell, 1994).

Formation of peptide-loaded HLA molecules (also called MHC I in other vertebrates) is a complex process and involves many steps (Bouvier, 2003): Newly synthesized MHC I α chains are translocated to the ER and associate with Calnexin (Degen and Williams, 1991). After binding of β_2-microglobulin, Calnexin dissociates and the MHC I chains form the so-called peptide-loading complex (Murphy, 2017), consisting of TAP, the oxidoreductase ERp57, calreticulin and tapasin (Blees et al., 2015). Through a process called peptide editing, weakly bound peptides on HLA molecules are exchanged for peptides with higher affinity (Fisette et al., 2016). Binding of peptides with sufficiently high affinity leads to disassembly of the peptide-loading complex and the loaded HLA molecules are ready for transport to the cell surface. Figure 1.2 shows a model of the peptide-loading complex.

The genes encoding for most HLA molecules are very polymorphic (see section 1.2) and differ in their binding affinity to different peptides. Peptides loaded onto HLA molecules typically have a length of about 8-11 amino acids (Rist et al., 2013), but peptides up to 33 residues have been observed (Stryhn et al., 2000). Peptide binding to HLA often depends on anchor residues (Grey et al., 1995), and computational software exists that predicts HLA allele specific binding affinity (O'Donnell et al., 2020; Reynisson et al., 2020).

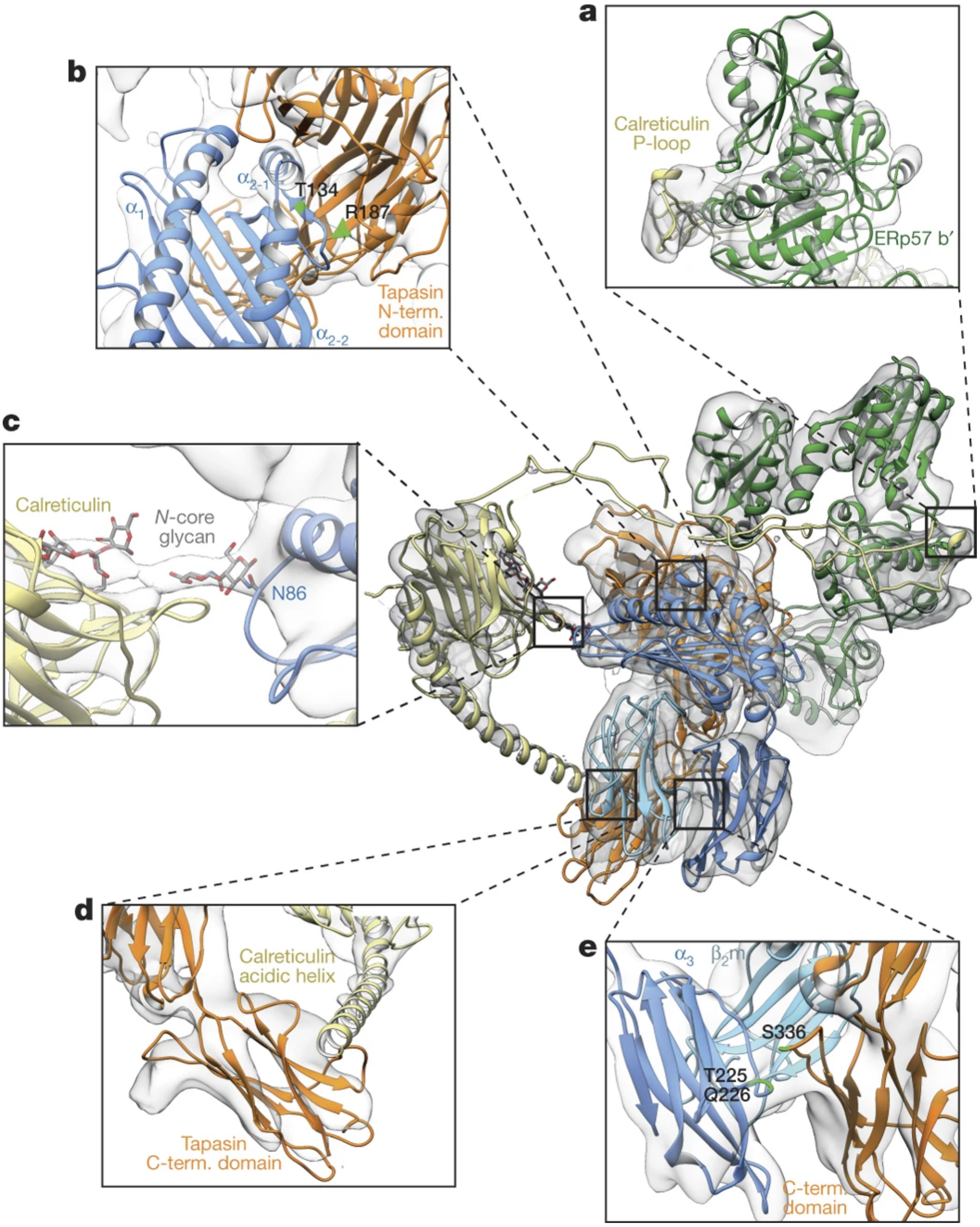


Figure 1.2: The model of the PLC editing module docked into the cryo-EM density (center) highlights important interactions between the ER chaperone network and the MHC-I client. a-e, Interactions of calreticulin and ERp57 (a), MHC-I heavy chains and tapasin (b), calreticulin and MHC-I heavy chains (N-core glycan) (c), calreticulin (acidic helix) and tapasin (C-terminal domain) (d), and tapasin (C-terminal domain) and MHC-I heavy chains (α_3) (e).

Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Structure of the human MHC-I peptide-loading complex, Blees, A., Januliene, D., Hofmann, T. et al. (2017)

1.1.3 Transport of the HLA/peptide complex to the cell surface

As other transmembrane surface proteins, peptide-loaded HLA molecules bud as COPII-coated vesicles from the ER membrane and are transported to the ER/Golgi intermediate compartment (ERGIC) and then via the Golgi apparatus to the cell surface (Adiko et al., 2015; Barlowe and Miller, 2013). The C-terminal amino acids valine or alanine act as export signals (Cho et al., 2010) and HLA molecules with suboptimally bound peptides are recycled between the ER and Golgi (Hsu et al., 1991). Inside the Golgi apparatus, glycans on the surface are modified (Wolfert and Boons, 2013). The glycosylation pattern has been shown to be important for immune function (Ryan and Cobb, 2012). Following glycan modifications at the Golgi apparatus, transport to the cell surface is facilitated by CD99 (Sohn et al., 2001). Once on the cell surface, HLA molecules can be recycled through endosomal internalization (Montealegre and van Endert, 2019).

The macro- and microeconomics of antigen presentation are complex: Measured on L929 cells, to maintain 2.6×10^9 proteins, each cell's 6×10^6 ribosomes produce around 4×10^6 proteins per minute. The total number of proteasomes in a cell is around 8×10^5, each degrading about 2.5 substrates per minute. Focusing on a single peptide, degradation of 2000 substrates on average is required to result in a single HLA molecule with that peptide on the cell surface (Princiotta et al., 2003).

In total, the reported median number of HLA molecules per cell ranges from about 5000 to 150000 (Schuster et al., 2017) and up to 3 million (Lanoix et al., 2018), allowing for up to 30000 different peptides to be presented on the cell surface (Kuznetsov et al., 2020).

1.2 Molecular structure of HLA molecules and nomenclature of HLA alleles

HLA class I molecules are heterodimers consisting of a heavy chain (α) and a light chain called β_2 microglobulin (Li et al., 2016). The α chain consists of three domains, an immunoglobulin-like domain near the cell membrane and two other domains that together form the peptide binding groove. Only the α chain is encoded by the MHC gene cluster and polymorphic. The β_2 microglobulin is non-covalently attached to the α chain and is not involved in peptide-binding and also does not have a transmembrane domain (Bjorkman et al., 1987). Figure 1.3 shows the 3-dimensional structure of an HLA class I molecule consisting of the α chain (blue), the non-covalently bound β_2 microglobulin (orange), and a bound peptide (red).

The molecular interactions that facilitate peptide binding are complex (Rammensee et al., 1993b): The peptides are held in place by the free carboxy- and amino termini at both ends of the peptide. Synthetic peptides lacking these terminal groups fail to bind in a stable manner to HLA molecules (Bouvier and Wiley, 1994; Murphy, 2017).

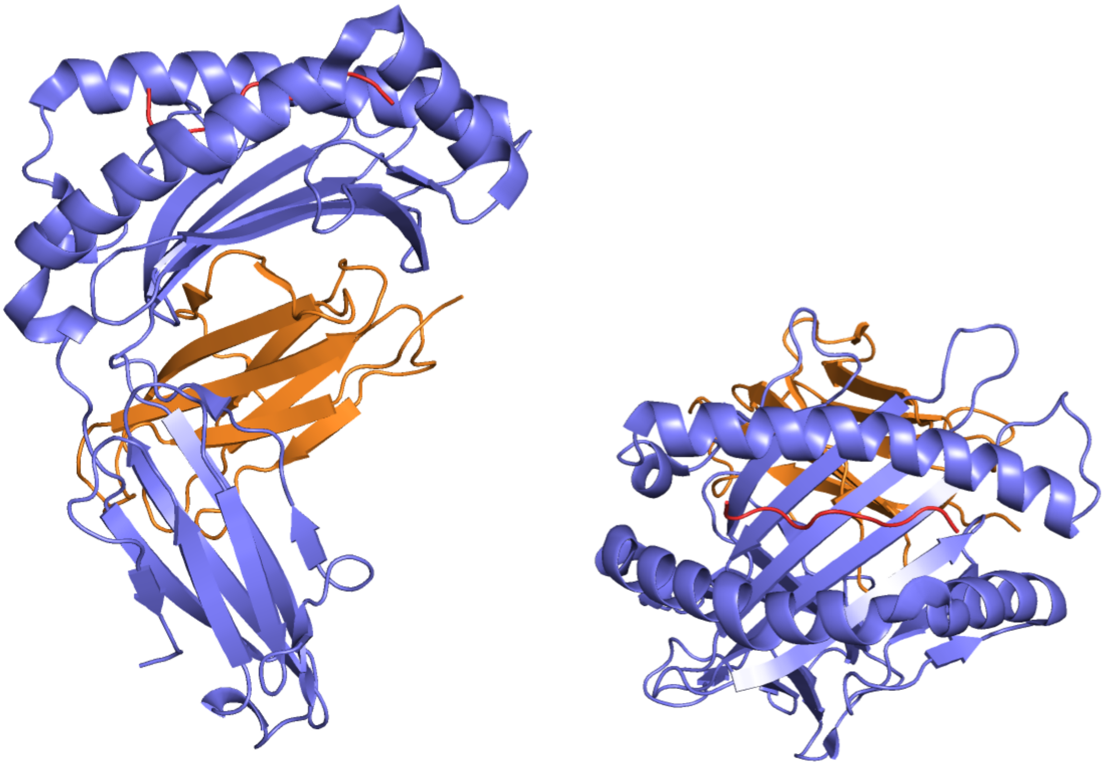


Figure 1.3: Peptide-loaded MHC class I molecule. The α chain is shown in blue, β_2 microglobulin is shown in orange, the bound peptide is shown in red. The image on the left shows a view from the side, the image on the right shows a view onto the peptide binding groove. PDB: 6P23 (Li et al., 2019)

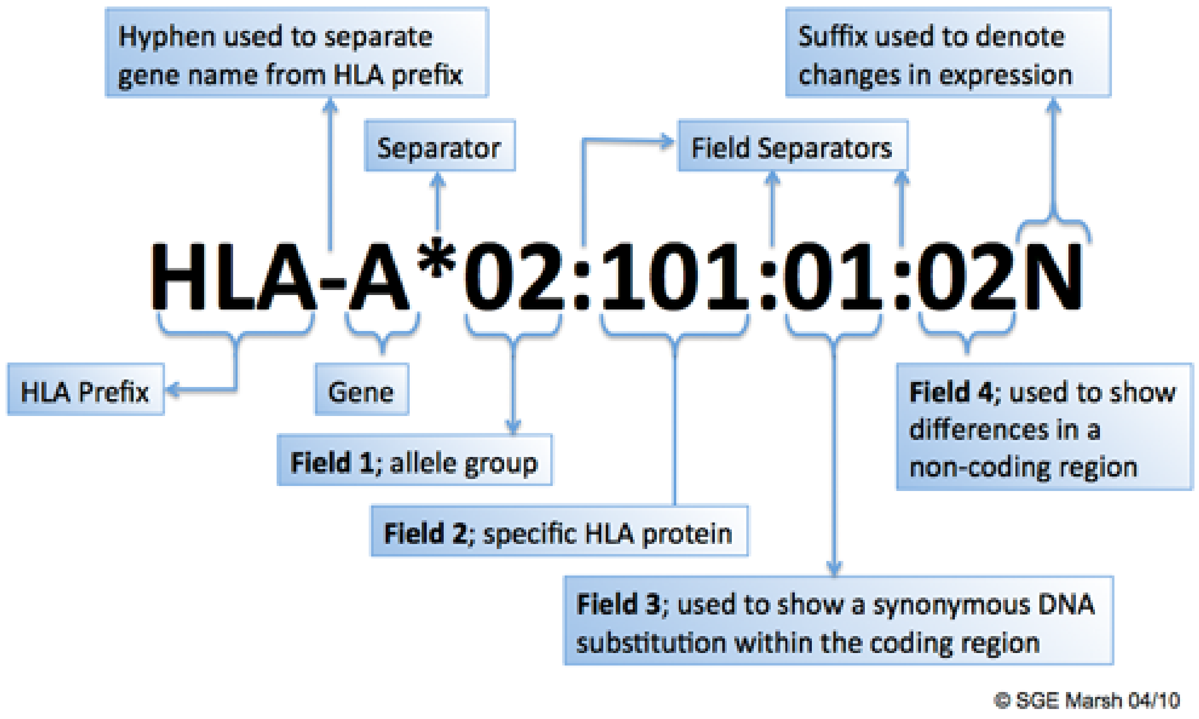

Hyphen used to separate gene name from HLA prefix
Separator
Field Separators
Suffix used to denote changes in expression
HLA-A*02:101:01:02N
HLA Prefix
Gene
Field 1; allele group
Field 2; specific HLA protein
Field 3; used to show a synonymous DNA substitution within the coding region
Field 4; used to show differences in a non-coding region
© SGE Marsh 04/10

Figure 1.4: HLA nomenclature. Each HLA allele has a unique number denoted by up to four sets of digits. Image taken from http://hla.alleles.org/nomenclature/naming.html, accessed April 22, 2022.

Longer peptides are accommodated by bulging the peptide backbone or –less common– by a protrusion mechanism (Murphy, 2017; Stryhn et al., 2000). The C-terminal amino acid is usually hydrophobic or charged, and strong binding is typically facilitated by two anchor residues (Ibe et al., 1996; Rammensee et al., 1993a).

The MHC genes are the most polymorphic genes in the human genome (Reche and Reinherz, 2003; Robinson et al., 2017, 2019). In humans, there are three genes encoding for HLA class I α chains, called HLA-A, HLA-B, and HLA-C (The MHC sequencing consortium, 1999). The number of alleles for these genes is immense, with 6921 alleles for HLA-A, 8182 alleles for HLA-B, and 6779 alleles for HLA-C known as of June 2021 (Robinson et al., 2019). MHC genes are present in all (jawed) vertebrates (Klein et al., 1993) and are most likely developed by gene duplication from a common vertebrate ancestor (Nei and Rooney, 2005).

Historically, HLA molecules were serologically defined (Thorsby, 2009). Since 1968, HLA nomenclature is standardized by the Nomenclature for Factors of the HLA System (Marsh et al., 2010). HLA alleles follow the pattern HLA-X*Y:Z, where HLA- is the prefix designating HLA alleles, X is the HLA gene (e.g. A-, B- or C for classical HLA class I genes), Y is the allele group (digits roughly corresponding to serotypically defined HLA antigens) and Z are digits to specify the specific HLA allele. Further fields separated by a colon may denote synonymous DNA substitutions in the coding region, DNA substitutions in the non-coding region and a suffix may be used to denote changes in expression. An example of a valid HLA allele according to this nomenclature is HLA-A*02:110:01:02N, Figure 1.4 shows a graphical representation of the HLA nomenclature.

1.3 Development of cytotoxic T cells

T cells originate from a common lymphopoietic precursor in the bone marrow (Heinzel et al., 2007; Serwold et al., 2009). These precursor cells migrate to the thymus, guided by –among other factors– P-selectins and CCL25 (Gossens et al., 2009; Zlotoff and Bhandoola, 2011), where they undergo a series of selection steps (Klein et al., 2014), see figure 1.5.

Histologically, the thymus consists of a cortex and medulla region. Using standard hematoxylin and eosin staining, the cortex appears darker, as it is the region of cell proliferation and early differentiation, whereas a series of negative selection steps occurs primarily in the medulla (Pearse, 2006).

T cell development has been most extensively studied in mice. Large differences between the T cell development in mice and humans exist (Kumar et al., 2018; Mestas and Hughes, 2004): For example, humans are born with a fully functional complement of T cells (Burt, 2013), whereas mice are born lymphopenic (Min et al., 2003). Similarly, the pool of naive T cells in mice is almost exclusively replenished from thymic output, whereas the majority of human naive T cells is derived from peripheral T cell division (den Braber et al., 2012).

Upon arrival in the thymus, commitment of the progenitor cells to the T cell lineage is mediated by Notch1 signaling (Pui et al., 1999). These immature thymocytes undergo a series of differentiation steps. These steps are characterized by the expression of different receptors and surface molecules, most importantly CD3, CD4, CD8, CD25 and CD44 (Godfrey et al., 1993). Thymocytes at the beginning of this differentiation process are called double-negative, because they lack both CD4 and CD8, as well as CD3 (Fowlkes et al., 1985). Double negative thymocytes can give rise to two important T cell lineages: The α:β T cells (Nikolić-Žugić, 1991), called after their T cell receptor consisting of an α and a β chain, and γ:δ T cells (Adams et al., 2015), a minority T cell lineage which neither expresses CD4 nor CD8 (Lew et al., 1986). In contrast to α:β T cells, γ:δ T cells have innate immune functions (Beetz et al., 2008).

Similar to B cells, somatic VDJ recombination is responsible for the vast T cell receptor repertoire (Schatz et al., 1992). In a first step, D_β and J_β genes are being recombined. In a following step V_β genes are concatenated to the DJ_β fragment, resulting in a complete gene of the β chain of the T cell receptor (Bassing et al., 2002). This newly formed β chain is expressed together with a pre-α chain, forming the pre-T cell receptor (pTCR) (von Boehmer and Fehling, 1997). Intracellular signaling eventually leads to expression of the CD8 and CD4 receptors (Yamasaki et al., 2005), which act as co-receptors and bind to constant regions of MHC class I and MHC class II molecules, respectively (Li et al.,

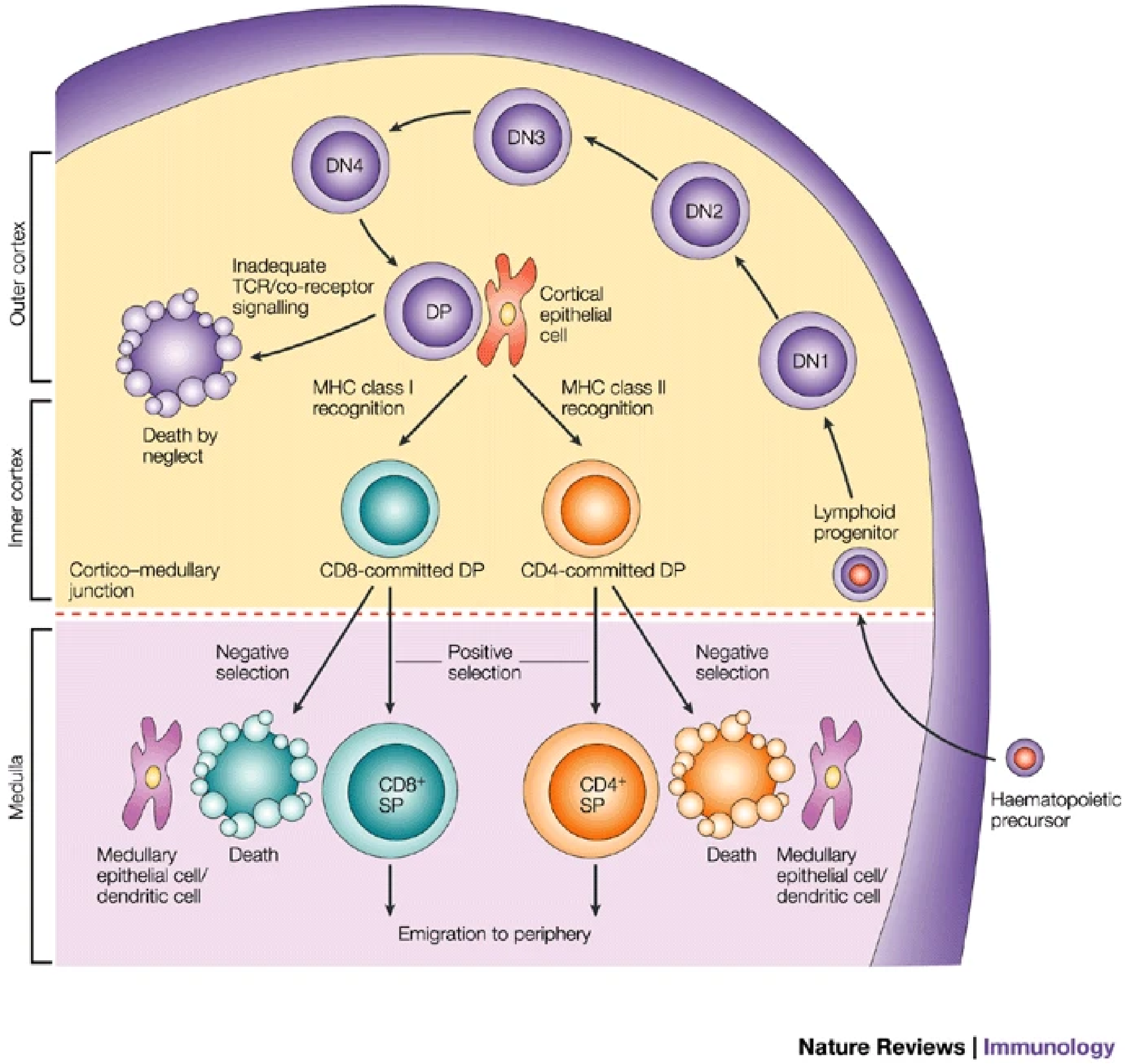


Figure 1.5: Overall scheme of T-cell development in the thymus. Hematopoietic precursor cells migrate from the bone marrow into the thymus, where they undergo a series of differentiation and selection steps. Thymocytes at the beginning of these differentiation steps are called double-negative (DN), because they lack both CD4 and CD8. During maturation, the T-cell receptor is formed through VDJ-recombination. CD4 and CD8 is expressed, and the thymocytes are now called double-positive (DP). These double-positive cells undergo both positive- and negative selection: Cells failing to bind either MHC class I or MHC class II undergo apoptosis, whereas cells that do bind to MHC class I or MHC class II emigrate to the periphery after further maturation steps.

Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, T-cell development and the CD4–CD8 lineage decision, Germain, R. (2002). Figure caption modified from the original.

2013).

After a series of proliferation steps, the VJ genes of the α chain genes start recombination to form a fully functional T cell receptor (Huang and Kanagawa, 2001). In contrast to B cells, rearrangement is not a one-time process, but can occur until successful rearrangement has taken place (Petrie et al., 1993).

In this double-positive stage, positive selection and negative selection occurs (Klein et al., 2014; Takaba and Takayanagi, 2017). To ensure that mature T cells can recognize antigen, double positive thymocytes are positively selected, where cells failing to bind to either MHC class I or MHC class II molecules undergo apoptosis (Surh and Sprent, 1994). This process has shown to enhance repertoire recognition of foreign antigens (Mandl et al., 2013). Negative selection also occurs during or after the double positive stage. In this step, cells binding too strongly to MHC are destroyed, as they might elicit strong autoimmune responses (Enouz et al., 2012; Klein et al., 2014).

1.4 Effector functions of cytotoxic T cells

T cells with a matching T cell receptor are able to bind to foreign epitopes and trigger cytotoxic effector functions. Because accidental activation of cytotoxic T cells can have devastating consequences (Liblau et al., 2002), this process is tightly regulated (Goronzy and Weyand, 2008; Mustelin and Taskén, 2003): Binding of the TCR and CD8 to MHC class I molecules is usually not enough to activate cytotoxic T cells (Bretscher and Cohn, 1970; Mueller et al., 1989), successful activation is dependent on co-stimulatory signals from CD4$^+$ T helper cells and professional antigen-presenting cells like dendritic cells (Feau et al., 2012), for example via the release of IL-2 (Luckheeram et al., 2012; Ross and Cantrell, 2018) and CD28 ligation (H Sepulveda and Dutton, 1999). Binding of the TCR to MHC class I molecules without co-stimulation leads to T cell anergy (Schwartz, 2003).

Upon T cell activation, an intracellular signaling cascade (Smith-Garvin et al., 2009) is triggered through phosphorylation of immunoreceptor tyrosine-based activation motifs (ITAMs) on the TCR-associated protein-complex CD3 (Samelson et al., 1986) via the tyrosine-protein kinase Lck (Barber et al., 1989).

Ligand-binding induces a series of rapid expansion (Curtsinger et al., 2003). The resulting effector T cells have heterogeneous phenotypes and differ in their effector functions (Han et al., 2011). Effector mechanisms are also broad, ranging from the release of inflammatory cytokines like IFN-γ and TNF-α (Bhat et al., 2017; Brehm et al., 2005), ligand-mediated apoptosis through Fas/FasL (Kagi et al., 1994) and release of pro-apoptotic molecules like perforins and granzymes (Trapani and Smyth, 2002). The time-span between TCR/MHC I binding and cell death ranges from 6 to 30 hours (Bhat et al., 2014).

After pathogen clearance, most of the effector T cells die through apoptosis, but a small fraction of about 5% differentiates into memory T cells (Cui and Kaech, 2010; Williams and Bevan, 2007), an important constituent of T cells-based adaptive immunity.

Chapter 2

Bayesian Data Analysis

The publications in section 3 rely on a branch of statistics called Bayesian statistics, whose defining property is the description of model parameters using probability distributions. I will introduce the main idea behind Bayesian statistics first, before contrasting it to classical (hereafter called frequentist) statistics, which is a branch of statistics most commonly taught in statistics courses. In the following sections, I will highlight the Bayesian workflow, with special consideration to model testing. The following sections will also introduce Pareto-smoothed importance sampling, a method used to estimate model performance on unseen data without refitting the model, and average predictive comparisons, which can be used to make the output of some statistical models more easily interpretable.

2.1 Bayesian statistics

2.1.1 The notion of probability

Before describing Bayesian statistics in further detail, it might be helpful to take some moment to reflect what probability is. Perhaps surprisingly, there is no universal "correct" definition of probability, the notion of probability depends on the context and can mean different things. "From a purely abstract, mathematical perspective, probability is simply a positive conserved quantity that can be distributed across a given space – in particular it does not necessarily refer to anything inherently random or uncertain." (Betancourt, 2018). This is the so-called *axiomatic interpretation* of probability, famously outlined by Kolmogorov (1933), who laid the foundations of modern probability theory.

However, when statisticians talk about probability, they usually do not refer to the abstract mathematical definition, but use the term to describe uncertain events. The perhaps easiest of these more practical interpretations of probability is the *Laplacian* definition of probability, which defines the probability of an event as the ratio of the number of favorable cases to the number of all cases possible (Laplace, 1814). The scope of this interpretation of probability is quite limited: It requires that the number of all possible outcomes can be enumerated, and that they are equally likely (otherwise we run into a circular argument).

Another interpretation of probability, that also covers events in which the number of all possible outcomes is not known in advance, is the so-called *frequentist* definition of probability, which interprets probability as the frequency of occurrences in a (hypothetically) infinite number of random trials (Venn, 1888). But what about events that do not fit into the framework of repeated random trials? Can we reason about the probability of the earth getting destroyed by an asteroid? This is related to the famous *sunrise problem*, introduced by Laplace (1814) to show limitations of inductive reasoning. In cases like this, it might be helpful to adapt a *Bayesian* interpretation of probability, that is more loose and focuses on using probability (in the mathematical sense) as a way to quantify uncertainty (Cox, 1946; de Finetti, 1974; Savage, 1972).

It should be noted that the concepts of the Bayesian and frequentist interpretation of *probability* do not directly map to Bayesian and frequentist *inference*, e.g. we frequently employ frequentist interpretations of probability in Bayesian models when talking about calibration or when viewing parameter priors or posteriors as giving information about what is known about the “true” parameter values (Gelman and Hennig, 2017). It is important to keep in mind that these distinctions are not merely philosophical, but have fundamental implications on the way statistical models are built in practice. In particular, viewing model parameters as fixed but unknown quantities (corresponding to the frequentist interpretation of probability) leads to convenient theoretical guarantees (Neyman and Pearson, 1933), whereas viewing model parameters as uncertain quantities that are described probabilistically allows for a workflow of repeated cycles of model inference, model critique and model improvement (Gelman et al., 2020).

The Bayesian interpretation of probability is sometimes criticized as “subjective”, in the sense that, compared to the frequentist interpretation, it does not “allow locating probabilities in an objective world that exists independently of the observer” (Gelman and Hennig, 2017). This point is addressed in greater detail in the section 2.4.

2.1.2 Bayes’ theorem

Bayesian inference is named after Thomas Bayes, an English mathematician and pastor. His work “An essay towards solving a problem in the doctrine of chances.” was published posthum in 1763 and describes a special case of what is today known as Bayes’ rule or Bayes’ theorem.

Bayes’ theorem is a way to invert conditional probabilities: Consider two random variables A and B. The probability of B, given that A occurred can be written as $p(B|A)$ (”probability of B given A”). If one were interested in the probability of A given B instead, it is possible to use Bayes’ rule to invert this conditional probability:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} \tag{2.1}$$

In introductory statistics textbooks, Bayes' rule is often introduced using the example of a diagnostic test. Assume a diagnostic test is performed with the following properties: If a person is affected by the tested disease, the test shows a correct positive result in 99% of the cases, when the person does not have the disease, the test shows a false-positive result in 5% of the cases. The disease is quite rare, with a frequency of 0.3% in the population. Given a positive test result, what is the probability that a person does have this disease?

The result to the above question can be calculated by plugging in the appropriate values into Bayes' theorem:

$$p(\text{ill}|\text{positive}) = \frac{p(\text{positive}|\text{ill})p(\text{ill})}{p(\text{positive})} \tag{2.2}$$

where $p(\text{positive}|\text{ill})$ is 99%, $p(\text{ill})$ is 0.3% and $p(\text{positive})$ can be calculated by combining the probability of true- and false positive results:

$$\begin{aligned} p(\text{positive}) = p(\text{positive}|\text{ill})p(\text{ill}) + p(\text{positive}|\text{healthy})p(\text{healthy}) = \\ 0.99 \times 0.003 + 0.05 \times 0.997 \approx 0.053 \end{aligned} \tag{2.3}$$

Plugging in all the values into the Bayes' rule gives:

$$p(\text{ill}|\text{positive}) = \frac{0.99 \times 0.003}{0.053} = 0.056 \tag{2.4}$$

A probability of about 5.6% given a positive test result might seem low at first, but can be understood in the context of the disease being rare compared to the false positive rate.

It is important to realize that the application of Bayes' rule in Bayesian statistics is not as narrow as presented here: Instead of applying Bayes' rule solely to event probabilities, its definition can be expanded to model parameters and observed data.

Consider being interested in the frequency of an HLA allele in a population. You are taking a random sample from the population of interest and perform genetic sequencing. Out of 100 samples, 20 persons have that allele. Denote the observed data as y, and the allele frequency as θ. Bayes' rule can then be written as:

$$p(\theta|y) = p(y|\theta)p(\theta)/p(y) \tag{2.5}$$

In Bayesian statistics, each individual component of Bayes' rule has a special name: $p(\theta|y)$ is called the posterior, because it reflects the information about the model parameter(s) conditioned on the observed data. $p(y|\theta)$ is called the likelihood and describes the probabilistic relationship between the observed data and the model parameters. Part of the process of model building is mapping the complexity of the real world to a simpler mathematical model, usually using probability distributions as building blocks. $p(y)$ is called the evidence and is the probability distribution of the data. It can be calculated as the integral of the likelihood over all possible parameter values. In most real-world applications, calculating this integral is computationally infeasible, so sampling based approaches are used instead that generate samples from the posterior distribution.

Therefore, the application of Bayes' rule in Bayesian statistics deviates in two important aspects from the application of Bayes' rule in most introductory statistics books:

1. It is not limited to simple event probabilities, but is applied to data and model parameters.
2. The plugged in values are not limited to probabilities, prior and likelihood are often probability distributions.

2.1.3 A primer in Bayesian modelling

Equation 2.5 also holds when θ is not a single parameter, but a collection of parameters (similar to how y is the whole observed data and not just a single data point). Consider the following example:

The efficacy of an antitumor treatment is studied in mice. In total, the sample consists of 10 mice in the control group, and 10 mice in the treatment group. After a set number of days, the mice are killed and the tumor mass is measured for each animal.

One way to quantify the treatment effect would be to model the tumor mass for each mouse as coming from a normal distribution with mean μ_i and standard deviation σ.

$$y_i \sim \text{Normal}(\mu_i, \sigma) \tag{2.6}$$

$$\mu_i = \alpha + \beta x_i \tag{2.7}$$

The mean μ_i is then modelled as the sum of two parameters: A general intercept α, which denotes the "baseline" tumor mass for both groups, and a treatment effect β, which denotes the mean tumor mass difference between the treatment and control group. x_i is a binary indicator variable with $x_i = 0$ if mouse i belongs to the control group, and $x_i = 1$ if mouse i belongs to the treatment group.

Equation 2.6 is the likelihood ($p(y|\theta)$), as it is a probabilistic description of the data, given the model parameters. What's missing for full Bayesian inference is the prior.

The prior can serve many purposes, for example to regularize estimates or induce sparsity. For the purpose of introducing Bayesian inference I settle for a weakly-informative prior, which can be thought of as a relatively safe default choice in a lot of situations. The goal of weakly-informative priors is to rule out unreasonably large estimates, but still allow for the estimates to be mostly governed by data. If we expect the mean tumor mass in the control group to be about 20 g, but unlikely 10 g or 30 g, a weakly informative prior on α could be $\alpha \sim \text{Normal}(20, 5)$. This is because the normal distribution allocates most (about 95%) of the probability mass within 2 standard deviations, so values above 30 g or below 10 g would be unlikely. Similarly, if we expect the treatment effect to be on the order of about 20% (e.g. from 20 g to 15 g), a weakly informative prior could be something like $\beta \sim$ Normal(0, 5). Note that we still center the prior for β on 0 to convey the expectation that we are skeptical about any possible treatment effect. For the standard deviation, we chose a positively constrained normal distribution with standard deviation 10, which allows for considerable differences in tumor mass between the animals.

$$\alpha \sim \text{Normal}(20, 5) \tag{2.8}$$

$$\beta \sim \text{Normal}(0, 5) \tag{2.9}$$

$$\sigma \sim \text{Normal}(0, 10) \tag{2.10}$$

The prior is the *joint* distribution of all the parameters, but in practice it is often convenient to define the prior in terms of marginals (that is, a distribution for α and a distribution for β), because it can be unwieldy to reason about the joint distribution of α and β.

This section is only intended as a short introduction, and although section 2.4 expands on the prior model a bit more, an in-depth treatment of prior modeling is given in Betancourt (2021).

Instead of relying purely on mathematical notation to set up a model, diagrams can be helpful to clarify the relationship between different parameters. In these diagrams, the observed data y is often shown at the bottom, with the dependency graph of the parameters and link functions shown above. Figure 2.1 shows a diagram of the model described in equations 2.6 to 2.10.

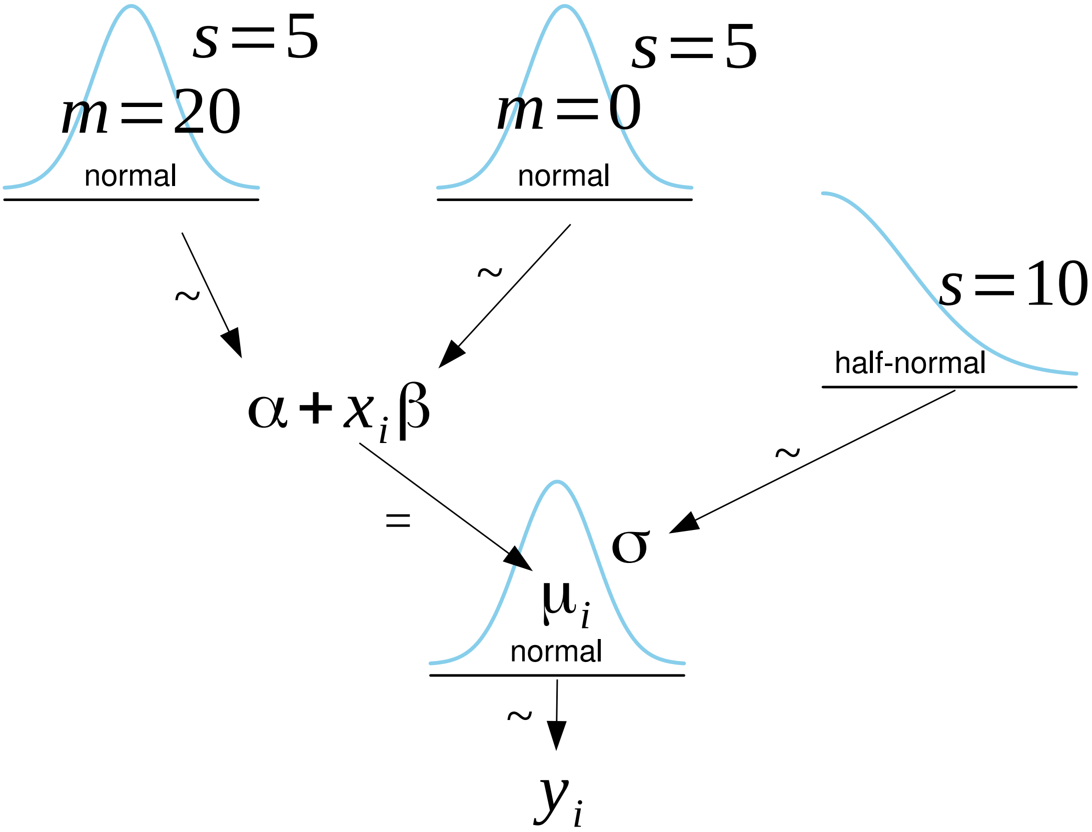


Figure 2.1: Diagram of the model described in equations 2.6 to 2.10. Reading from the bottom up, the observed data y is modelled as coming from a normal distribution with mean μ_i and standard deviation σ. The mean for each sample μ_i is calculated as the sum of an intercept α and a treatment effect β. In this model the intercept α can be interpreted as the mean tumor mass in the control group (because for the control group, $x_i = 0$ and therefore $\mu_i = \alpha$), and the treatment effect β can be interpreted as the difference in tumor mass between the control group and the treatment group. Diagram template from Bååth (2016).

2.2 Frequentist statistics

It is well possible to learn Bayesian statistics without a background in frequentist statistics. One might even argue that it is easier this way, since some concepts common in frequentist statistics do not apply in Bayesian settings.

An example for this is the inclusion of “non-significant” variables in a regression model. Common guidelines suggest excluding any non-significant variables in a regression model in order to reduce the standard error of the estimates. This goes against good practice in a Bayesian setting, because the fact that the data is consistent with no effect of the predictor on the outcome does not necessarily mean that it is not useful for prediction. In general, including additional predictors in a regression model does not harm inference in a Bayesian setting, because all correlations between parameters are captured in the posterior.

In a perhaps more extreme example, some authors even advocate against checking model assumptions, because "using the data twice" might lead to inflated type I errors (Rochon et al., 2012). On the other hand, model checking is a cornerstone of the Bayesian workflow and necessary for building good statistical models (Gelman et al., 2020).

Nevertheless, it might be beneficial to also briefly introduce frequentist statistics, simply because this is what is taught in most statistics classes and what most people applying statistics are used to. Being able to build upon something may help to better understand the differences in Bayesian and frequentist thinking.

2.2.1 Null hypothesis significance testing

A statistical method which is often used to test scientific hypotheses is null hypothesis significance testing (NHST). In NHST, the researcher does not use the observed data to make inferences about the hypothesis of interest directly. Instead, a null hypothesis is formed, which typically conveys the assumption that none of the measured quantities have an effect on the outcome. This null hypothesis is compared to a (often favored) alternative hypothesis, that is assumed to be a better description of the data generating process. Both null hypothesis and alternative hypothesis are mathematical models that describe how the observed data could have been originated in a probabilistic fashion, where the alternative hypothesis is often the negation of the null hypothesis.

By calculating the probability of the observed data under the null hypothesis, the researcher tries to make deductions about the real world: If the probability of the observed data under the assumption of the null hypothesis being true is low, this is taken as evidence that the null hypothesis is not a good description of the real world, which is usually interpreted as evidence for the alternative hypothesis.

In this aspect, NHST is like a thought experiment: One imagines a state of the world as being true (the one formulated in the null hypothesis) and compares how compatible this state is with the observed data.

The probability of observing the observed (or more extreme) data, under the assumption of the null hypothesis being true is called the *p-value*.

2.2.2 Common mistakes in interpreting p-values

From the above description, common mistakes in interpreting p-values become apparent. In the following, common misinterpretations of p-values are listed, along with a description of why these statements are false. The p-value is **not**:

- *the probability of the null hypothesis being true.*
  The p-value is the probability of observing the observed *data* under the assumption that the null hypothesis is correct. As it is the probability of observing data, it is not the probability of the null hypothesis being true.

- *a statement about the favored, alternative hypothesis.*
  As the p-value is strictly conditioned on the null hypothesis being true, it does not make statements about other hypotheses. For example, the mathematical model behind the data generating process could be a poor description of the real world. In this case, low p-values would occur even in the absence of any effect. Similarly, there might be several possible alternative hypotheses, and rejection of the null model does not necessarily mean that the favored alternative has to be correct.

- *the probability that the observed result has occurred by chance.*
  The p-value is calculated under the assumption that the null hypothesis is correct. However, even in the absence of any treatment effect the underlying mathematical model can still be a poor description of the real world. Often, there are also several possible phrasings of the idea of “no effect”, which means that the mathematical model underlying the null hypothesis might just be one of many and other equally reasonable explanations would result in different p-values.

- *a summary for the magnitude of an effect.*
  The p-value is only a statement about the compatibility of the observed data with the null hypothesis. Being incompatible with the null hypothesis does not mean that the difference between groups has to be large: For example even small, biologically irrelevant differences might result in very low p-values if the data-generating process is not a good description of the data. Even if the null model is a good description of the real world, low p-values can still occur when the data is large. In this case, even small, meaningless deviations from the null model can result in extremely low p-values.

2.2.3 Issues with p-values and null hypothesis significance testing

Aside from these issues of interpreting p-values, there are also several, more fundamental issues when using null hypothesis significance testing as a method to gain insight about the real world: The formal view of the p-value as a probability conditional on the null is mathematically correct but typically irrelevant to research goals (Gelman, 2013). Researchers are usually not concerned about the probability of observing certain data under the assumption of some hypothesis being true, but want to gain insight into how strong the hypothesis is supported by the observed data. In mathematical notation, the former can be expressed as $p(D|H)$, that is the probability of data given the hypothesis, whereas what is usually desired is $p(H|D)$, that is, a statement about the hypothesis of interest, given the data (conditioned on a mathematical model).

Other issues exist: An important property of p-values is often ignored, which is that they are dependent on the data collection process. As they are the probability of observing the observed data, it is crucially important how the sampling procedure is framed into a mathematical model (in statistical terms called the sampling distribution). The p-value for evaluating the null hypothesis of "a coin is unbiased" when the stopping criterion is to throw the coin 50 times is different from the p-value obtained by the stopping criterion "throw until the coin showed heads 25 times". It is easy to construct examples in which the p-value under one data collection procedure is significant (using a typical threshold of 0.05), while being far away from that threshold under another data collection procedure, even if the observed data is identical in both cases.

The dependence on p-values on the data collection procedure has important implications on the scientific process: Consider a researcher who, after collecting data, obtains a p-value of 0.1. Still being unsure about the scientific relevance of this finding, the researcher then decides to collect more data. The eventuality of collecting data was not accounted for in the first analysis, so the obtained p-values lose all meaning. The effect of optional stopping on statistical inference through p-values is described in Pocock (1983), and the author shows that with just two additional data collections, the probability of obtaining statistical significance when the null hypothesis is in fact true in their example is 11% (instead of 5% using a significance threshold of 0.05). The point is not that researchers might cheat the system to achieve statistical significance, but that collecting more evidence in light of unclear data is such an integral part of science that statistical procedures should easily allow that, even when not initially planned.

Statistical issues of using p-values for scientific questions are well described in the literature, for selected examples along with some opposing views see: Altman and Bland (1995); Gelman (2013); Gelman and Stern (2006); Greenland (2019); Ioannidis (2005).

2.3 Advantages and disadvantages of Bayesian statistics

The interpretation of probability as a means to express partial knowledge about a system, as opposed to the definition of probability based on repeated sampling, has several advantages:

- *Bayesian statistics provides a full, probabilistic description of all model parameters, along with their correlation and uncertainty*
  This does not only provide more information about the main quantity of interest in many data analyses, but also provides an extremely useful way of model checking:

By repeatedly sampling from the likelihood function for different model parameter values, it is possible to generate replicated data sets from a probabilistic model. In other words, instead of going from the usual direction of data to parameter values, the information about the distribution of model parameters can be used to generate new data under the model's assumptions. This method is called *posterior predictive check* (Gabry et al., 2019). By comparing these simulated data to the actual observed data it is possible to implicitly check the assumptions of a statistical model. This has the advantage over explicit checks of being more sensitive and exhaustive: Statistical models make many assumptions, some of which cannot be quantified easily. By carefully analyzing in which ways replicated data deviates from the observed data, it is possible to gradually improve a model until it is consistent with the observed data in the degree that is necessary for the model to be useful.

- *The results of Bayesian inference are not dependent on the stopping criterion of the data collection procedure*
  As the Bayesian definition of probability does not rely on hypothetical repeated sampling, the results only depend on the actual observed data and not on hypothetical unobserved data. This makes it possible that a researcher might collect data, determines that the gathered evidence is not sufficient and gathers more data later on. While frequentist statistics allows for data collection procedures like this if they are specified in advance, this is seldom the case in real world scenarios. It is important to note that this is only true when the decision criterion is the width of the posterior distribution. Using other stopping criteria, e.g. the tail probability of a parameter being larger/smaller than 0, still biases estimates towards that criterion (Deng et al., 2016; Kruschke, 2013).

- *Bayesian statistics allows for the principled integration of prior information.*
  Prior information is an integral part of science and might come in the form of prior studies or additional sources of information, e.g. a mutation in a viral genome is more likely to be an HLA escape mutation if we know that an epitope around that position in the multiple sequence alignment is restricted by that particular HLA allele. Making use of prior information does not always mean to gather additional data, prior information almost always exist in the form of physical constrains and general knowledge about the problem structure. For example, the proportion of lymphocytes in a tissue sample of a tumor is unlikely to exceed 50%, or the transmission speed of a neuron will not exceed the speed of light. While seemingly unimportant, ruling out these extreme values in a statistical model provides a useful way to get more stable estimates and greatly reduces the probability of overestimating effect sizes (type M error) or estimating the direction of an effect in the wrong direction (type S error) (Gelman and Carlin, 2014).

- *In Bayesian statistics it is easy to include data containing uncertain measurements*

or missing information
Not only model parameters can be quantified using probability distributions, it is also possible to treat data as uncertain, e.g. because of measurement error or completely missing data values. In this case, the uncertainty of the measurements directly translate into uncertainty of the estimates.

- *Bayesian statistics focuses on quantification of uncertainty.* Bayesian statistics places a strong focus on quantifying the magnitude of an effect. A probability distribution conveys more (relevant) information than a point estimate with confidence intervals. Compared to NHST, "how much and how uncertain" is usually a better data summary than a binary rejection/non-rejection of a null hypothesis. There is the possible objection that NHST also provides point estimates along with confidence intervals, but these are not statements about inferential uncertainty, but bounds defined under the assumption of repeated sampling. Therefore, a Bayesian posterior distribution, summarized as a 95% posterior interval, usually provides more relevant information.

Despite these advantages, Bayesian statistics is also confronted with a series of drawbacks, which are the focus of current research efforts:

- *Bayesian inference relies on approximate Markov chain Monte Carlo methods and is often subjected to numerical difficulties.*
  In most real world scenarios, closed-form solutions of a Bayesian posterior distribution do not exist. Therefore, Bayesian inference relies on Markov chain Monte Carlo methods, which generate *samples* from the posterior distribution. The drawback of these methods is that it is sometimes difficult to generate representative samples from the posterior distribution due to numerical difficulties. This can occur for example due to strongly correlated parameters or tight "ridges" in the posterior geometry. For these reasons, diagnostic checks have been developed that help to diagnose sampling related issues. Nevertheless, a significant proportion of time spent building Bayesian models is working around sampling related issues.

- *Running large models may be computationally expensive.*
  As Bayesian inference relies on gradient based methods, obtaining samples from the posterior involves repeated gradient calculations, which can be computationally expensive for models with many parameters. Additionally, Markov Chain Monte Carlo is hard to parallelize, therefore it is not straight forward to tackle complex models with more computing power.

- *Bayesian statistics is a quickly moving field.*
  Many algorithms and computational tools have only been developed in recent years, and best practices have not been formed conclusively. This makes it difficult to keep up for people who are primarily interested in using statistics as a research tool.

2.4 Bayes and Subjectivity

The need to specify a prior distribution over the model's parameters is sometimes met with unease, out of concern for making a statistical analysis become subjective (Gelman and Hennig, 2017). Indeed, one of the main objectives of science is to learn something about the objective reality we live in, so what place do apparently "subjective" procedures have? Wouldn't it be possible to choose a prior distribution in such a way that any analysis could support any research hypothesis? I believe this fear is unfounded, for several reasons:

The prior is not the only model component where a researcher has to make conscious decisions

The guiding principle of statistical modeling is to map relevant aspects of reality onto a simpler model. The simpler model allows drawing conclusions that can then be mapped back to the real world. By design, this means that a statistical model is a (non-reversible) abstraction of the real world and many possible such mappings exist. Any statistical procedure therefore necessarily has to make some assumptions, e.g. which variables to include in a model, how to handle outliers or how to model missing data. The prior is not special in that regard.

Adding prior information can help to better describe the observed data

Bayesian and frequentist methods alike are often likelihood-based, that is, they formulate a probabilistic description of the data given the model parameters. And while it is true that the prior can only be understood in the context of the likelihood (Gelman et al., 2017), the likelihood can also be understood as being embedded into the proper context through the prior. Or more clearly: The prior restricts the effect of the likelihood to a region of reasonable parameter values and therefore lets it match objective reality more closely by removing obvious mismatches between the model and the real world. I would argue that a model that includes prior information is actually more objective than a model which does not, simply because it is a better description of the real world. Prior information does not always have to mean expert knowledge, even information about the order of magnitude of measurements can be useful information. Under these considerations, I think it is unjustified to hold the prior to a different standard than the likelihood. One might object that in contrast to the prior, the choice of a particular likelihood function can be validated in the limit of infinite data, but arguing based on the hypothetical existence of infinite data does not seem like a strong counterargument to the observation that both the prior and the likelihood can be somewhat arbitrary.

A prior does not need to be subjective

When methods are being labelled as subjective or objective, these terms generally do not follow a precise definition. Most often, the word subjective is used to describe methods that rely on information not included in the data.

Bayesian statistics can follow different schools of thought. A “subjective” interpretation of Bayesian statistics follows the idea of de Finetti (1974) that probabilities are based on personal belief and can be quantified by considering how a rational decision maker would bet on the occurrence of an event. On the other extreme, “objective Bayesians” try to set up objective prior distributions by focusing on the idea that the prior should convey no information. A popular example of such a prior is Jeffrey’s prior (Jeffreys, 1946), which defines a prior in such a way that it is invariant to all transformations of a parameter.

However, a prior does not necessarily have to be based on the concept of “no information”. The popular R-package rstanarm (Goodrich et al., 2020) uses priors that depend on the data by default, e.g. for a regression model, the prior on the regression coefficients is $\beta_k \sim \text{Normal}(0, 2.5 s_y / s_x)$, where s_y is the standard deviation of the outcome and s_x is the standard deviation of the predictor x_k. This prior is not subjective in the sense that it is not based on any information but the observed data.

Objective methods often only appear objective

One may object that such a rule is arbitrary, but so are estimators in frequentist statistics: Estimators in frequentist statistics usually follow the principle of unbiased minimum variance. These are certainly reasonable properties, but the unbiasedness does not come for free, as these estimators typically have higher variance than methods that introduce a small bias. From a practical perspective, being right on average (this is what it means to be unbiased) does not have to be best if all we have is a single dataset. This can be seen from the recent increase in popularity of regularization methods in frequentist statistics, like Lasso (Santosa and Symes, 1986; Tibshirani, 1996) or ridge regression (Hoerl and Kennard, 1970). These methods produce biased estimates in the sense that estimates are pulled towards 0, which is a desirable feature in many applications. Their popularity shows that estimators do not necessarily have to be unbiased, and that the criterion of unbiased minimum variance is just one of many possible trade-offs.

A similar trade-off is the focus on the type I error in null hypothesis significance testing. It is possible to show that certain hypotheses tests are optimal (Robinson, 1979), that is, they provide the minimum possible type II error (failing to reject a null hypothesis that is actually false) while controlling for the type I error (rejecting a null hypothesis that is actually true). These methods are objective in the sense that they are optimal when the objective is to minimize the type II error while controlling for the type I error, but this requires that (just) controlling the type I error is desired in the first place. When effect sizes are small and the variance between measurements is large, like it is often the case in Biology, statistical significant results can often be in the wrong direction and greatly overestimate an effect (Gelman and Carlin, 2014). A researcher is definitely interested producing fewer estimates that are in the wrong direction, so the narrow focus on statistical significance is again just one of many possible trade-offs.

The example of unbiased estimators and type I errors in hypothesis tests share a common problem: they try to decouple decision analysis and statistical analysis; or framed differently, they imply certain preferences of decision-making. This is what makes them appear to be objective: Under the implied cost-benefit profile, they are optimal decision rules. Unfortunately, these implied preferences rarely match the actual preferences. This means that it does not make sense to completely decouple a statistical analysis from the decision-making process behind it. If a certain drug is known to have little to no side effects, it might make sense to administer it even if the effectiveness is not shown conclusively.

There is no statistics police

One might agree with the stance that it is not useful to label statistical methods as subjective or objective, but that it is still helpful to follow a generally agreed upon set of default tools that provide a "level playing field" for scientists, so that a scientist who wishes to publish a certain finding cannot simply choose the method according to the desired result. This view considers statistics as a "gatekeeper" to protect against scientific misconduct. Methods should therefore allow for as little individual impact on the result as possible.

There are two important issues with this view: The first is that even if we would impose strict rules on which statistical methods to use, researcher's degrees of freedom would still be an issue. In fact, even when using statistical significance as a filter, flexibility in data collection and analysis allows presenting anything as significant (Simmons et al., 2011). It is simply impossible to impose such strict rules that a malicious actor would not be able to find plausible evidence for any research hypothesis. Statistics therefore cannot function as a gatekeeper to protect against scientific misconduct.

The second issue is that the idea of a single hypothesis that is being tested goes against scientific reality. Research hypotheses rarely arise in a vacuum, they are the result of previous hypotheses and explorative data analysis. If we view science as more of an iterative process, adhering to strict rules therefore hides important elements of scientific learning.

Guiding principles

Once we realize that objectivity is a weak criterion by which to select a statistical method, is it time to despair and lose all hope? How do we make choices in data analysis when all of them seem arbitrary? A starting point may be to reflect why "objective" methods are so appealing: they limit individual impact. If a statistical procedure does not have any tuning parameters, it is not possible to set them to the wrong value. If we limit individual impact, it makes it less likely for a reviewer to disagree with our decisions. In some aspects, this is the statistical manifestation of *choice overload* (Iyengar and Lepper, 2000), a (controversial, (Chernev et al., 2015)) phenomenon in psychology.

I argue that one important aspect for the drive towards objectivity is the circumstance that many statistical methods offer little additional information beyond the inference, they are not testable. If the analysis is a black box, it makes sense to prefer methods that have as few knobs to turn as possible. This leads to the first principle that one could follow when choosing between several statistical methods: Use methods that provide feedback on how well they worked for a specific analysis. This naturally leads to preferring Bayesian models, because treating model parameters as probabilistic helps tremendously with model checking.

A list of "guiding principles" I follow in my own work looks like this:

- *Prefer methods that are testable.* When we view statistical models as abstractions of the real world, a useful model requirement is to explain the observed data. If a model is unable to explain the observed data in some relevant aspect, this means that the abstraction was too coarse, and we did not capture some important aspect of the real world. While this is only a necessary (and not a sufficient) condition, requiring that a statistical model explains the observed data already drastically reduces the model space.

  This does not mean that a model that does not describe the observed data well cannot be useful, one could easily imagine a case were a model does not explain the observed data but works fine on future data. However, there should also exist a model that works well on observed *and* future data, so restricting ourselves to models that do explain the observed data can be interpreted as a safety measure against those models that work poorly on both observed and future data.

  One possible concern when selecting models in this fashion is overfitting. When iteratively searching for models that work well on the observed data, our design decisions are driven by the observed data and models might perform poorly on unseen, future data. This point is discussed in section 2.5.

- *Fail fast.* Prefer methods that are sensitive to misspecifications. For example when comparing samples on the group level, it might make sense to build the model on the layer of individual observations and aggregate the individual effects to group effects later on, instead of estimating the group effects directly. This helps to make model misspecifications become apparent more easily, for example because some individuals systematically differ from others in the same group. This point is particularly relevant for hypothesis testing, because the probability of the observed data is often evaluated by calculating test statistics that summarize data sets by a single number. These reduced models have the advantage of being widely applicable, but the drawback is that it is difficult to test them and find unexpected patterns in the observed data.

- *Build models iteratively.* Iterative model building is a cornerstone of Bayesian model building. This point is addressed in section 2.5.

- *Include as much relevant information as possible.* Occam's razor tells us to prefer simple explanations. But when building statistical models, simplicity is often seen as a disadvantage. For example, we often have to argue why certain relevant predictors were not included, why we did not model non-response when analyzing a survey or why we did not adjust for confounding variables. So why does statistics seemingly go against the scientific ideal of simplicity? I believe this is a matter of perspective: When we view models as abstraction machines that capture some aspect of actuality, a "simple" model is a model that abstracts less. Unfortunately, the world is complex and we understand little, so abstraction is a necessary requirement for a model to be useful.

  So how do we find the Goldilocks model, one that provides enough abstraction, so we can understand what is happening, but is close enough to reality, so we can learn something? A good starting point is usually to consider the information we want to make use of and how we can include that information in a statistical model. As an example, consider HAMdetector (section 3.1). When presented with the task of identifying HLA-associated mutations, information can come in different ways: The obvious source of information are the sequence counts, e.g. how often we observe a replacement in hosts with and without that respective allele. But we also have access to sequence data, so it is possible to infer a phylogenetic tree. Additionally, we have (implicit) information of epitope binding affinity from epitope prediction tools.

  So how can we piece all of this information together? We expect mutations within the boundary of a certain epitope to be more likely HLA-associated, so we use a prior that allows us to model the expected degree of sparsity (the number of HLA-associated mutations compared to the total number of alignment positions). The phylogenetic tree tells us something about the relationship between the sequences, so we adapt the likelihood to not treat all sequences as independent. In this way, each piece of additional information helps us to come up with a suited model structure.

  Including as much information as possible also helps with model testing. The more we know, the easier it is to systematically test for discrepancies between the model's expectation and our knowledge. This makes it more difficult to build a model that is consistent with all information, but also increases our confidence in the model if we do find a model that is consistent with all of our domain knowledge.

- *Prefer mechanistic models, sometimes even when you just care about prediction.* Statistical models have many use cases: Sometimes, a model is built to provide predictions about the future. At other times, the goal of a model is to interpret the inferred model parameters and learn something about the underlying process. Of course, these goals are not completely separate, sometimes we are interested in a model that predicts as well and yields useful parameter estimates.

When the main goal is prediction, and we are lucky to have a lot of data, a popular class of models are neural networks. This kind of machine learning approach is attractive because these models are able to learn the model structure from the data. For example when confronted with the task of identifying neutrophils in microscopic images, we don't have to explicitly include the properties of neutrophils (e.g. segmented nuclei) in the model, it is enough to provide the model with examples of how a neutrophil looks like. This approach is not without downsides however, because they lack in interpretation, e.g. after fitting the model, it is difficult to learn how neutrophils differ from other cells. The issue of interpretability is an active area of research in the machine learning community.

There is another downside however, which is the required amount of data. Learning the structure of a neural network is a difficult task and requires many data points. Sometimes the advantages clearly outweigh the disadvantages, for example when we cannot specify the required model structure because we do not know which information to include or because we cannot list them all. This is often the case in the field of computer vision. Including all properties of the objects a statistical model needs to identify would be a laborious task. It is also brittle because objects may look different from another angle, e.g. a cylinder is just a circle when viewed from above. Models like neural networks are sometimes called *non-parametric*, not because the models do not have parameters, but because the inferred parameters do not have a direct interpretation.

So in which situations should we choose non-parametric methods, and when should we use parametric approaches? The obvious situation in which we should use a parametric model is when we care about the parameter estimates. But specifying additional structure is also useful when all we care about is prediction: A model that includes important prior information is usually more precise than a model that does not have access to this piece of information.

When deciding between parametric methods, *mechanistic* models are often preferred over more general models. Mechanistic models are models that include some mechanistic understanding of how the observed data may have come about. A good example for this is the golf putting model in Gelman et al. (2020). When modelling the probability of success in golf putting as a function of distance from the hole, we could consider a logistic regression model of the form $y = \text{logistic}(a + bx_j)$, where y is the probability of success, x_j is the distance from the hole, logistic is the logistic link function and a and b are the model parameters.

This model does not include much information about the problem structure though, and indeed logistic regression fits the observed data quite poorly. Another way to model this data is to try to include some mechanistic knowledge about golfing: One can imagine a successful golf put as the result of two events: The golfer has to hit the ball at the correct angle, and the golfer has to hit the ball with a certain strength. A ball closer to the hole is easier to put than a ball further away because the range of angles that would lead to the ball going inside the hole is larger.

Mechanistic models are useful because they are rooted in domain knowledge. If we find such a model that fits the observed data, it is more likely for such a model to generalize to future data than for a model that does not include mechanistic information.

Another nice property of mechanistic models is that we can learn more from model misfits: If a mechanistic model provides a poor fit to the data, this definitely means that our understanding about the process of interest lacks some important piece of information. If including some piece of information improves the model fit, we can conclude that this additional input is an important factor for prediction. Parameters of mechanistic models are also often easier to interpret (and allow for more precise priors), because they are directly linked to the real world. For example just from thinking about the problem it becomes obvious that a golfer has to hit the ball within an interval of maximum 45∘, whereas the range of possible parameter values for a and b are not directly apparent.

- *Be skeptical by default.* It is usually beneficial to use methods that are conservative in their estimates, e.g. when constructing a prior with marginal distributions for each parameter, it might make sense to use distributions that are centered around 0 for the effect size parameters. This makes a statistical model more skeptical by default and helps to not overestimate possible effects.

- *Keep the folk theorem in mind.* The folk theorem of statistical computing (Gelman, 2008) states that problems with a statistical model often manifest as computational problems, for example as convergence issues when using Markov chain Monte Carlo methods. These computational problems often suggest changes in the model structure, e.g. re-parameterizing a model because of strong posterior correlations or including additional prior information.

- *Think about how the model is used later on.* Different models are suited for different applications and section 2.4 addressed the point that one drawback of strictly controlling the type I error is that it implies a certain cost-benefit trade-off which might not coincide with the actual intentions. Consider the area of preventive maintenance in industry: Large machinery relies on constant service of its parts and while servicing is expensive, a complete failure might have disastrous consequences. When building a statistical model to predict when to replace a component of a machine, one might be presented with the scenario that measuring some properties of that component requires shutting the machine off, whereas other properties can be continuously measured using sensors. A model that includes all predictors might work much better than a model that just relies on sensor data, but the cost-benefit trade-offs are drastically different. Which model to build depends on the specific trade-offs which can (and should) be included into the decision analysis.

2.5 Bayesian workflow

This section loosely follows the *Principled Bayesian Workflow* in Betancourt (2020).

The Bayesian workflow is iterative: Starting from an initial model, repeated cycles of model building, model evaluation and model improvement are used to come up with a model that is consistent with our domain expertise and can explain the observed data in all aspects we care about. This section gives an example of this workflow: Starting with the prior model, I introduce the concepts of *prior push-forward checks* and *prior predictive checks*

Up until this point, all steps in the Bayesian workflow are applied before collecting (or analyzing) any data. I then introduce Bayesian sample size calculation and experimental design, which are useful to determine which and how much data to collect.

The next steps focus on model evaluation: Posterior predictive checks are used to ensure that our model catches all relevant aspects of the observed data, and Pareto-smoothed importance sampling can be used to compare statistical models with each other. Once we are happy with the model, I briefly touch on general aspects of model visualization and introduction average predictive comparisons, a method that can help to interpret model parameters.

All these aspects of the Bayesian workflow are best explained with a specific example. The goal of this example is to be simple enough to not require any deep domain knowledge, while still being rich enough to illustrate all aspects of Bayesian modelling.

2.5.1 Example: Diabetes in dogs

The data presented in this example are entirely fictional and are not based on experimental data.

As in humans, the hormone insulin is also produced in dogs by beta cells in the pancreas. Insulin regulates the blood glucose level by increasing the intake of glucose of other cells of the body. If beta cells do not produce enough insulin, e.g. after being destroyed by an autoimmune reaction, the blood glucose levels cannot be sufficiently controlled and chronic hyperglycemia leads to the classic symptoms of diabetes: increased thirst, frequent urination, weight loss and increased hunger.

Diabetes in dogs can be treated in the same way as in humans: Injections of insulin, which has been extracted from pig pancreas or produced by recombination, lowers the blood glucose concentration and –together with dietary adjustments– help to treat diabetes. Unfortunately, administration of insulin does not always reduce the blood glucose concentration reliably and a switch to a different insulin product might be necessary.

Imagine you are developing a new insulin product to treat canine diabetes. Because your insulin is recombinant and much closer to the insulin naturally present in dogs, you expect it to have a greater efficacy than the competitor product, which is known to lower blood glucose levels in 50% of the treated dogs. A clinical relevant increase in efficacy over the competitor product is expected to be 5%.

2.5.2 Conceptual analysis

The first step of any statistical analysis is conceptual analysis. What are the relevant aspects our model needs to capture? What are the units of measurements? In this particular example, we care about the efficacy of the developed insulin product compared to the competitor product. The units of measurement are individual dogs, and the outcome is therapy success or failure. One could also imagine a continuous outcome, e.g. by measuring the reduction in blood glucose level after administration of the insulin. This is usually advantageous because it does not discard information (and we can always binarize the output later on), but for this example we assume that the output is therapy success or failure defined by some clinical criterion.

We then have to decide what measurements to include. Usually, this is constrained by the data we have access to or by financial considerations. However, thinking about what we measure and how it relates to the outcome of interest is a crucial step, because if our measurements are not connected to the outcome of interest, all of our modelling efforts will be in vain. For the purpose of this example, let's assume that we have access to veterinary records that include information about height (m), weight (kg), age (years) and breed.

In the next step we have to decide on how to relate the measurements to the outcome of interest for our initial model. This is of course an open-ended question and depends on personal preference, but the list in section 2.4 can give some guidance. For this example, we will stick with a logistic regression model, which is often a default choice when modelling binary outcomes. Our likelihood can be set up like this:

$$y_i \sim \text{Bernoulli}(\text{logistic}(\alpha + \beta_1 \times \text{height} + \beta_2 \times \text{weight} + \beta_3 \times \text{age})) \tag{2.11}$$

y_i is the outcome for dog i (0 for therapy failure, 1 for therapy success), logistic is the logistic link function, which is defined as $\text{logstic}(x) = exp(x)/(1 + exp(x))$, α is the intercept that can be interpreted as the baseline efficacy of the insulin, and β_1, β_2 and β_3 quantify the effect of our other predictors on the outcome.

Regression coefficients in logistic regression can be interpreted as expected changes in log-odds, e.g. a parameter value of 0.4 for β_2 would mean that on average, a dog that weighs 1 kg more than another dog with otherwise identical measurements has increased log-odds of therapy success of 0.4, where log-odds is the logarithm of the ratio p / (1 - p). Measuring regression coefficients on the scale of log-odds has the advantage that effects can be interpreted like other regression coefficients, e.g. 0 means no effect and negative values mean that the variable is negatively associated with the outcome of interest, but have the disadvantage of being a bit harder to interpret, i.e. it is not immediately clear if a change in log-odds of 1 correspond to a large or small effect. A helpful approximation is the so-called "divide by 4" rule, which says that a change in log-odds of x approximately correspond to a change of x/4 on the probability scale, e.g. a coefficient of 0.4 means that a 1 unit change in the input is associated with an increase in probability by 10%.

2.5.3 Prior model

In addition to the likelihood function, we also need a prior model for Bayesian inference, that is, we need a joint distribution of the parameters that then gets updated in the light of new data. As with the likelihood function (and all parts of modelling), there is no universal correct choice. But we do have some guidelines: The prior should be consistent with our domain knowledge. For example, we know that other insulin treatments are effective about 50% of the time. Considering that our insulin is recombinant and resembles the canine insulin more closely, we also expect it to be more effective. Of course, it could also be less effective, but probabilities near 0 and 1 are quite unlikely.

The process of coming up with prior distributions that accurately reflect domain expertise is called prior elicitation. In general, an effective and relatively easy way to elicit priors is by deciding on thresholds that separate unlikely from implausible parameter values. We require that a prior that is consistent with our knowledge only places little probability mass on parameter values that we deem implausible.

In our example, our domain knowledge is most easily described on the level of treatment success probabilities, but the model parameters are on the level of log odds. As it is generally difficult to reason about coefficients on a log-odds scale, one might be inclined to use broad priors that cover a wide range of possible parameter values. A possible prior could therefore be:

$$\beta_1, \beta_2, \beta_3, \beta_4 \sim \text{Normal}(0, 10)$$

Note that the prior is always a *joint* distribution of all model parameters, but in practice it is often difficult to define such a joint distribution directly. Even with just 4 parameters, we have to reason about a 4-dimensional space. It is therefore often convenient to assume prior independence between parameters and define the joint prior in terms of its marginal densities.

Our fully specified model can be described by the following diagram:

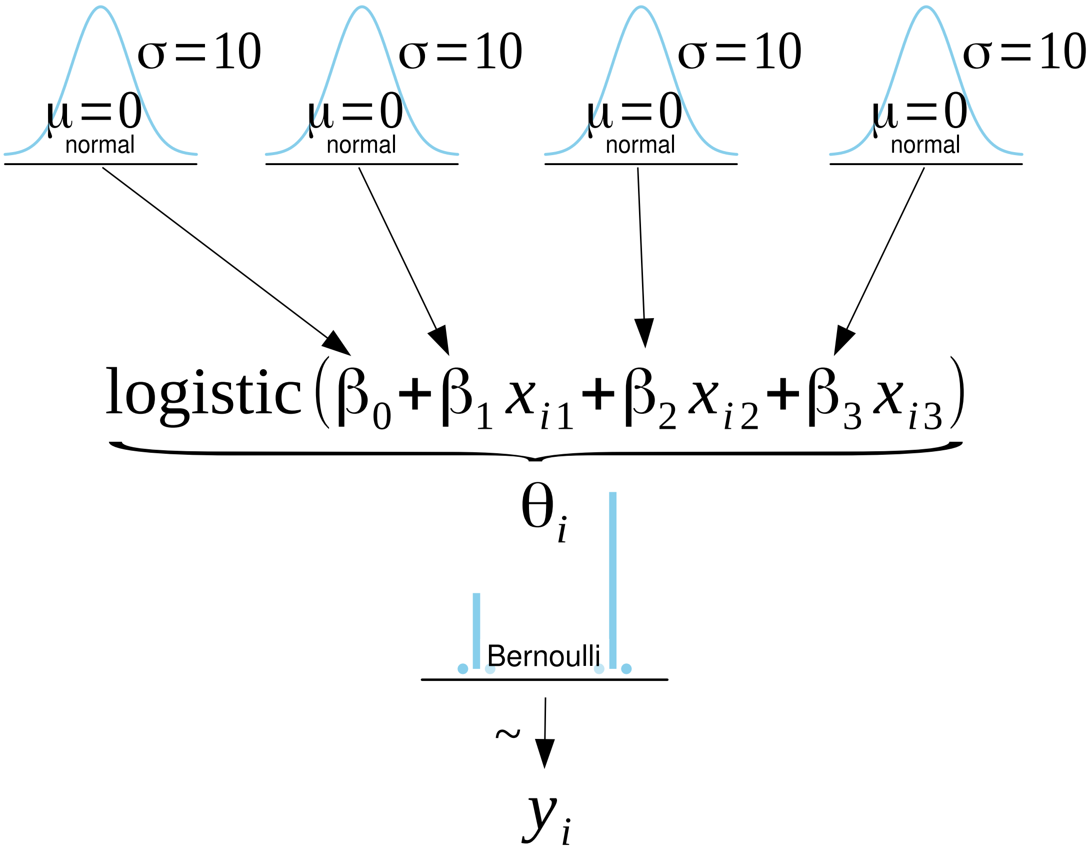


Figure 2.2: Diagram of the complete example model. Reading from the bottom up, the observed data y_i is modelled as coming from a Bernoulli distribution with success probability θ_i. The success probability is modelled with a regression term that is transformed to the interval [0, 1] with the inverse logistic function. Diagram template from Bååth (2016).

2.5.4 Prior push-forward checks

As stated previously, we have a good intuition about the expected treatment success probability θ_i, but it is difficult to reason about which treatment success probabilities we imply with our prior model. The implications of the prior model on transformed parameters can be investigated with *prior push-forward checks*, which push the prior probability density forward towards a summary function of our choice. This lets us explore if our prior model is consistent with what we know about different summaries of the model parameters. In this particular example, we can treat the estimated treatment success probability $theta_i$ as a summary function of our parameters: $\theta_i = \text{logistic}(\alpha + \beta_1 \times x_{i1} + \beta_2 \times x_{i2} + \beta_3 \times x_{i3})$.

We can easily simulate values of $p(\alpha, \beta_1, \beta_2, \beta_3)$, but the summary function also requires us to plug in some values of x_{i1}, x_{i2}, x_{i3}. If we already collected some data we could use the observed inputs as an approximation to the distribution of input values, but if we have not collected any data yet we can always investigate the push-forward distribution for "representative" data points. Figure 2.3 shows the expected treatment success probability for a dog with a height of 30 cm, a weight of 7 kg and an age of 5 years.

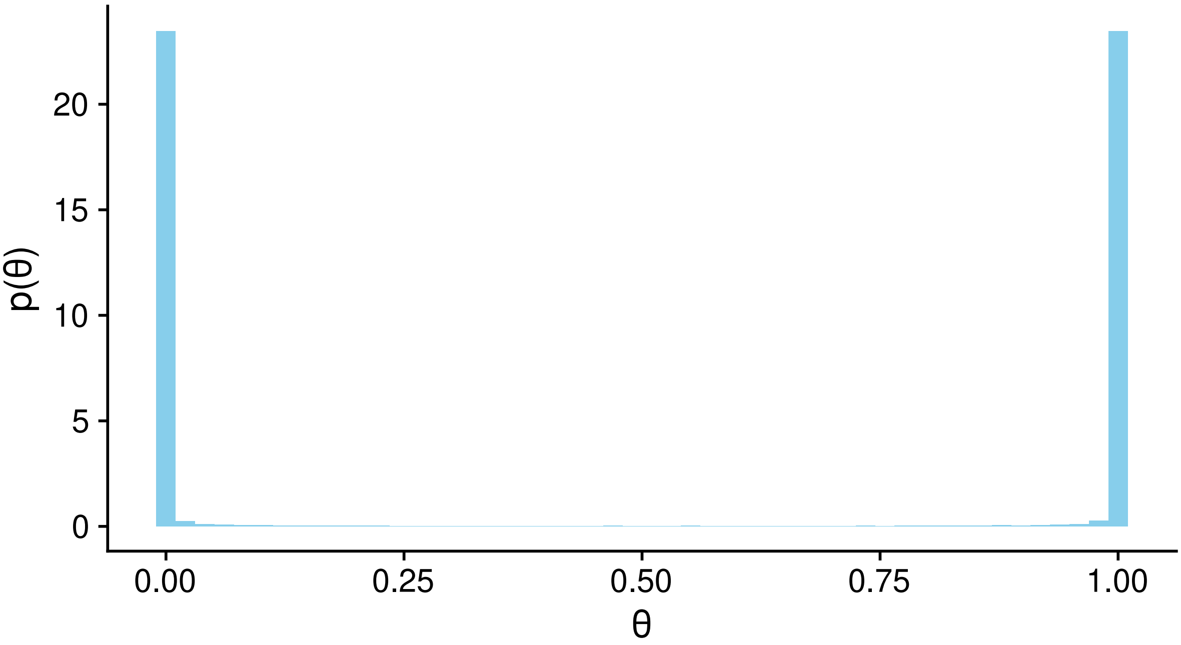


Figure 2.3: Implied distribution of treatment success probabilities for a dog with a height of 30 cm, a weight of 7 kg and an age of 5 years.

The plot shows that the implied prior on the treatment success probability is indeed in conflict with our domain knowledge, as it strongly concentrated around values of 0 and 1. The prior distribution that was supposed to be "weakly-informative" turns out to only allocate little probability mass at the region of 50%, where we expect the therapy success probability of our insulin to be.

This occurs because the regression term gets transformed by the logistic link function, which quickly reaches 0 or 1 for large negative and positive values, respectively (see Fig. 2.4).

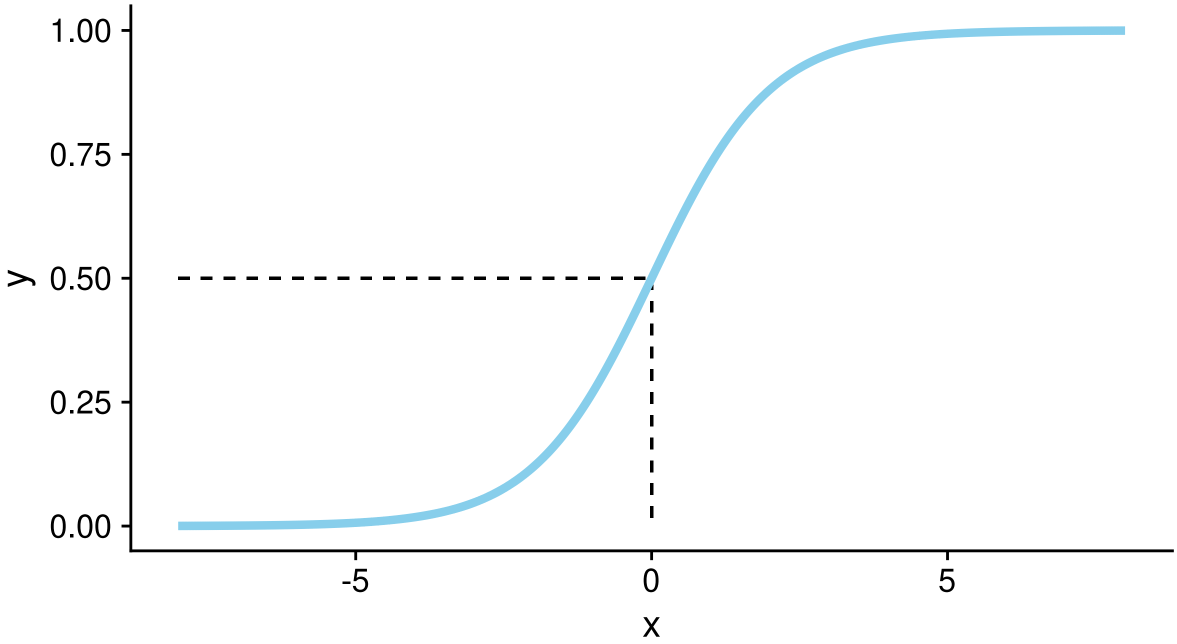


Figure 2.4: The logistic function $f(x) = \exp(x)/(1 + \exp(x))$ transforms inputs on the whole line of real numbers to an interval [0, 1].

We have to adapt the prior so that it is not in conflict with our domain knowledge anymore. Figure 2.4 shows that our inputs to the logistic function need to be closer towards 0 to allocate more probability mass in the region of around 50%.

We therefore change the marginal densities to:

$$\begin{aligned}
\alpha &\sim \text{Normal}(0, 1) \\
\beta_1 &\sim \text{Normal}(0, 1) \\
\beta_2 &\sim \text{Normal}(0, 0.2) \\
\beta_3 &\sim \text{Normal}(0, 0.1)
\end{aligned}$$

One might be hesitant to use such a prior because it seems to be too narrow, but when applying the “divide by 4” rule we can see that such a prior is still consistent with very large effect sizes: The normal distribution allocates most probability mass within two standard deviations from the mean, the marginal prior on β_1 is consistent with parameter estimates up to around 2. While a height difference of 1 m is a large difference for dogs, the divide by 4 rule tells us that a change in height by 1 meter can correspond to a change in treatment success probability by 50% (2/4) in either direction.

For weight, a difference of 1 kg can correspond to a change in treatment success probability by up to 10% and for age, a difference of 1 year can correspond to a change in treatment success probability by up to 5%. The updated prior push-forward distribution is shown in Figure 2.5

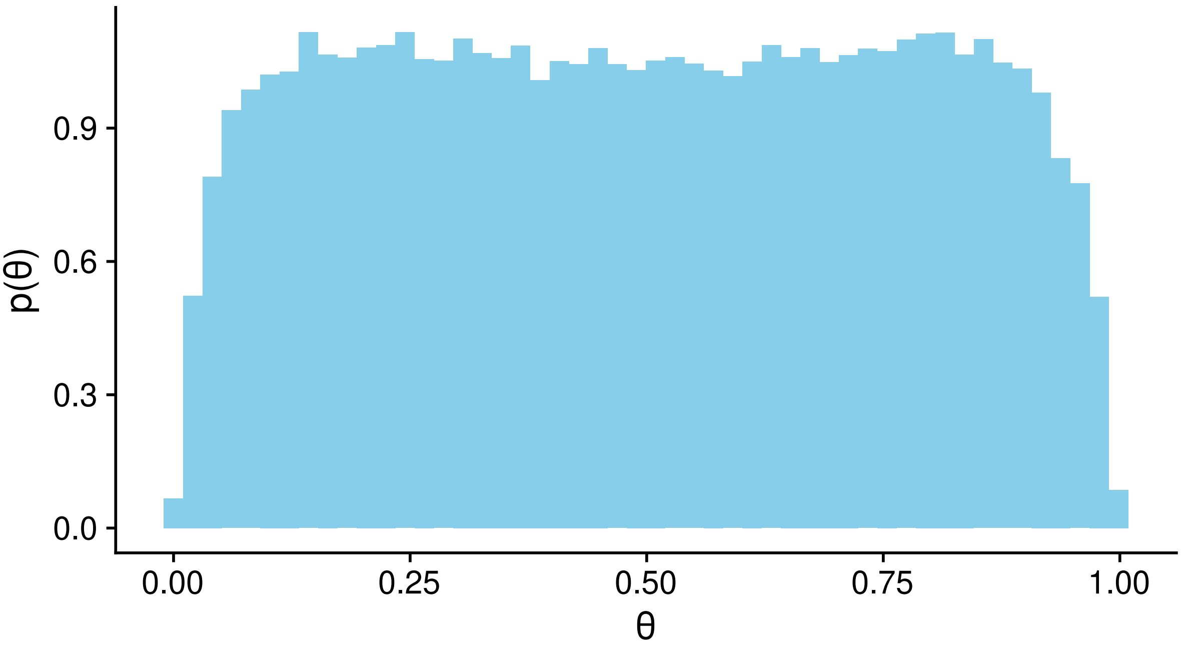


Figure 2.5: Updated implied distribution of treatment success probabilities for a dog with a height of 30 cm, a weight of 7 kg and an age of 5 years.

This implied distribution of treatment success probabilities is not in conflict with our domain knowledge anymore, as it allocated enough probability mass to treatment success probabilities around 0.5.

2.5.5 Prior predictive checks

Prior push-forward checks are used to study the implications of the prior on summary functions of the model parameters, but we can also go one step further and check the implications on the prior predictive distribution itself. This is helpful because it also includes the likelihood function in the model check. Looking at the prior predictive distribution is particularly useful for continuous observations. Consider for a moment that we had measured blood glucose concentrations instead of a binary treatment outcome. Our prior model would imply a distribution of changes in blood glucose levels, and we could compare this distribution with our domain knowledge. For example, if our prior model would imply blood glucose concentrations that are not in agreement with living dogs, that would mean we would have to update our prior. We can also compute arbitrary summary functions, e.g. compare 99% quantiles with expected extreme values.

2.5.6 Sample size determination

Bayesian model building is an iterative process, which is best applied *before* data has been collected. When data collection is expensive and time-consuming, as it is often the case in fields like biology and chemistry, simulated data can be used to determine the number of samples required to achieve parameter estimates of the desired accuracy. The required sample size may depend on several factors: In general, models that include problem-specific structure need less data than models that do not take this information into account. Additionally, prior knowledge can help to reduce the necessary amount of data. An often overlooked aspect is the quality of measurements: If the measured variables are not well-connected to the outcome of interest, no amount of data can help to yield useful estimates. Effect size is another important contributing factor, a low signal-to-noise ratio means that fewer data is required to get estimates that are precise enough.

Simulations might reveal that the budget is not large enough to yield useful estimates and can therefore help to reduce costs by not conducting an experiment that would not be useful. The simulations might also help to avoid using too many samples, which in addition to the economical benefit is particularly import for animal-based research.

Simulating data under a hypothetical model requires assumptions about the distribution of the data and a (preliminary) probabilistic model that links the observed data to the quantities of interest. Simulations are still useful however, because a.) the main goal of these simulations is to get a rough idea about the required amount of data, b.) the preliminary models are a useful foundation for the model building step and c.) simulations can be run for different candidate models and candidate data sets, so that the uncertainty about these assumptions translates into uncertainty about the required sample size.

A frequentist approach to determine sample sizes is called power calculation, which determines the required sample size by assuming an effect of a specific magnitude and then calculating the sample size required to obtain statistical significance in x% of hypothetical data collections, where x is a large number, usually 80%. These power calculations can also be applied in a Bayesian context, i.e. how many samples are required so that a posterior interval excludes 0 with a specified probability. Here, we are interested in the required number of samples to achieve estimates with a certain precision. To achieve this goal, we employ the following strategy:

1. Simulate observed inputs, i.e. height, weight and age values of dogs.

2. Draw a set of model parameter values from the prior distribution.

3. Simulate a set of observations based on the model parameters.

4. Fit the model.

5. Draw samples from the posterior predictive distribution.

6. Calculate the width of the distribution of treatment success probabilities based on the posterior predictive distribution.

Figure 2.6 shows the resulting required sample size to obtain 80% posterior intervals based on quantiles with a given interval width.

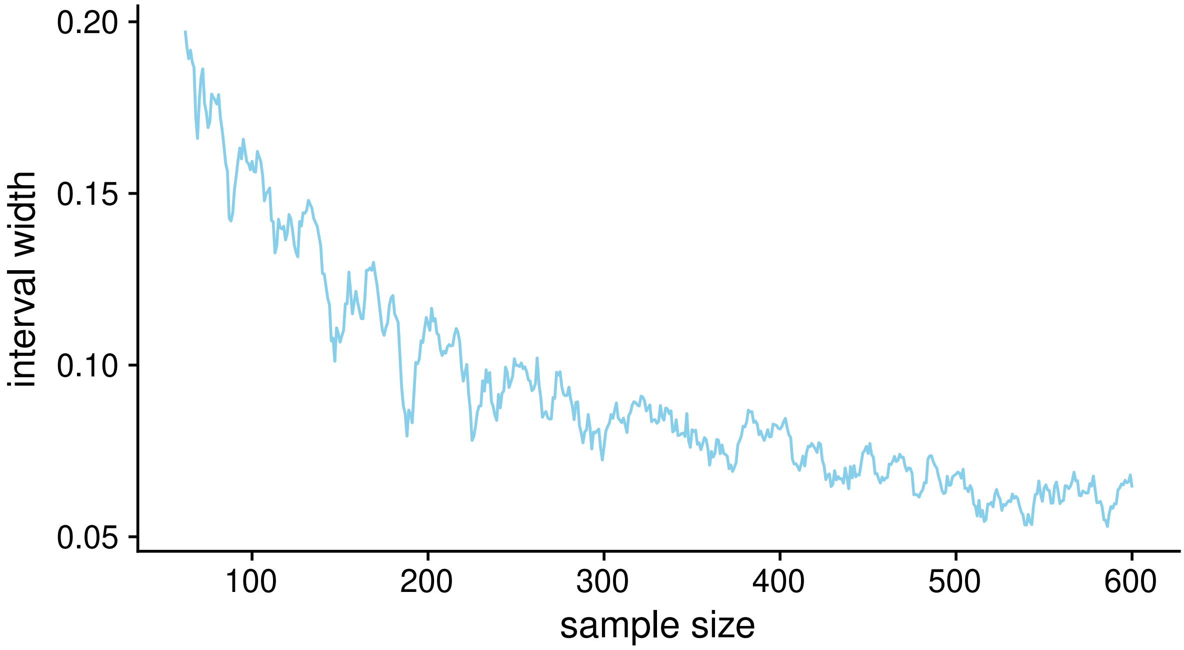


Figure 2.6: Width of 90% posterior intervals based on quantiles for a given sample size. Dog heights were simulated from $x_1 \sim \text{Uniform}(0.3, 1.2)$, weights were simulated from $x_2 \sim x_1/4 + \text{Normal}(0, 1)$, age values were simulated from $x_3 \sim \text{Uniform}(1, 15)$ (rounded to the nearest integer). The plot shows that in order to achieve posterior intervals with a width of about 10%, a sample size of about 300 is needed.

2.5.7 Posterior predictive checks

One of the main aspects of the Bayesian workflow is model testing. Bayesian models belong to a class of models called generative models, which means that they can be run “backwards”, i.e. instead of going from data to estimated parameters, parameter estimates can be used to generate new data under the model’s assumptions. This allows for an extremely powerful form of model testing: By comparing data generated by the model to the actual observed data, shortcomings of the model can be identified. This method is called *posterior predictive check*. The goal is to find data summaries that highlight important aspects of the data: For example when working on a model that accesses failure risks of machine parts, getting the tails of the observations right is extremely important. A useful data summary could then be the 1% and 99% quantiles. A histogram of 99% quantiles of data sets generated under the model’s assumption could be compared to the actual observed quantile. If the 99% quantile of the observed data does not lie within the range of 99% quantiles expected under the model's assumptions, this highlights a potential issue for drawing conclusions based on the model results.

Consider we collected the height, weight, age and breed of 400 dogs and now want to compare the observed treatment outcomes to the treatment outcomes expected under the model's assumptions. One simple summary could be the observed proportion of treatment successes compared to the proportion of treatment successes of simulated data from the posterior predictive distribution. Figure 2.7 shows a histogram of the proportion of treatment successes from the draws of the posterior predictive distribution with a line showing the true proportion of treatment successes observed in the data. The line lies well within the region of the histogram, showing no deviations between the expected and observed data.

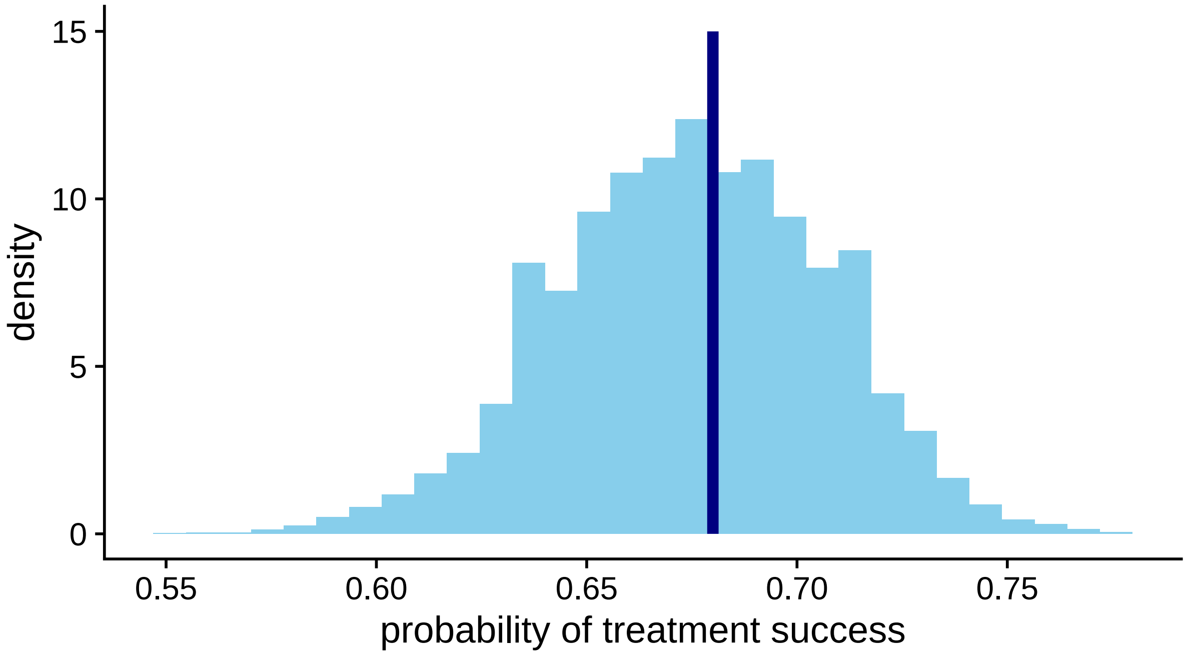


Figure 2.7: Posterior predictive check of the expected proportion of treatment successes under the model (histogram) and observed proportion of treatment successes in the data (vertical line).

Such a posterior predictive check does not provide much information however, because the summary statistic (the observed proportion of treatment successes) is quite insensitive towards model misfits. The Bernoulli likelihood function reaches its maximum around the observed mean of the data, therefore just about any logistic regression model should get the mean right. A more helpful posterior predictive check may be achieved by plotting different subsets of the data. Our model does not make use of information about the dog breeds (implicitly assuming that breed is not associated with treatment success). We can check if this assumption holds by comparing the total number of successfully treated dogs for each breed in the observed dataset compared to the expected total number of successfully treated dogs for each breed in draws from the posterior predictive distribution (Figure 2.8).

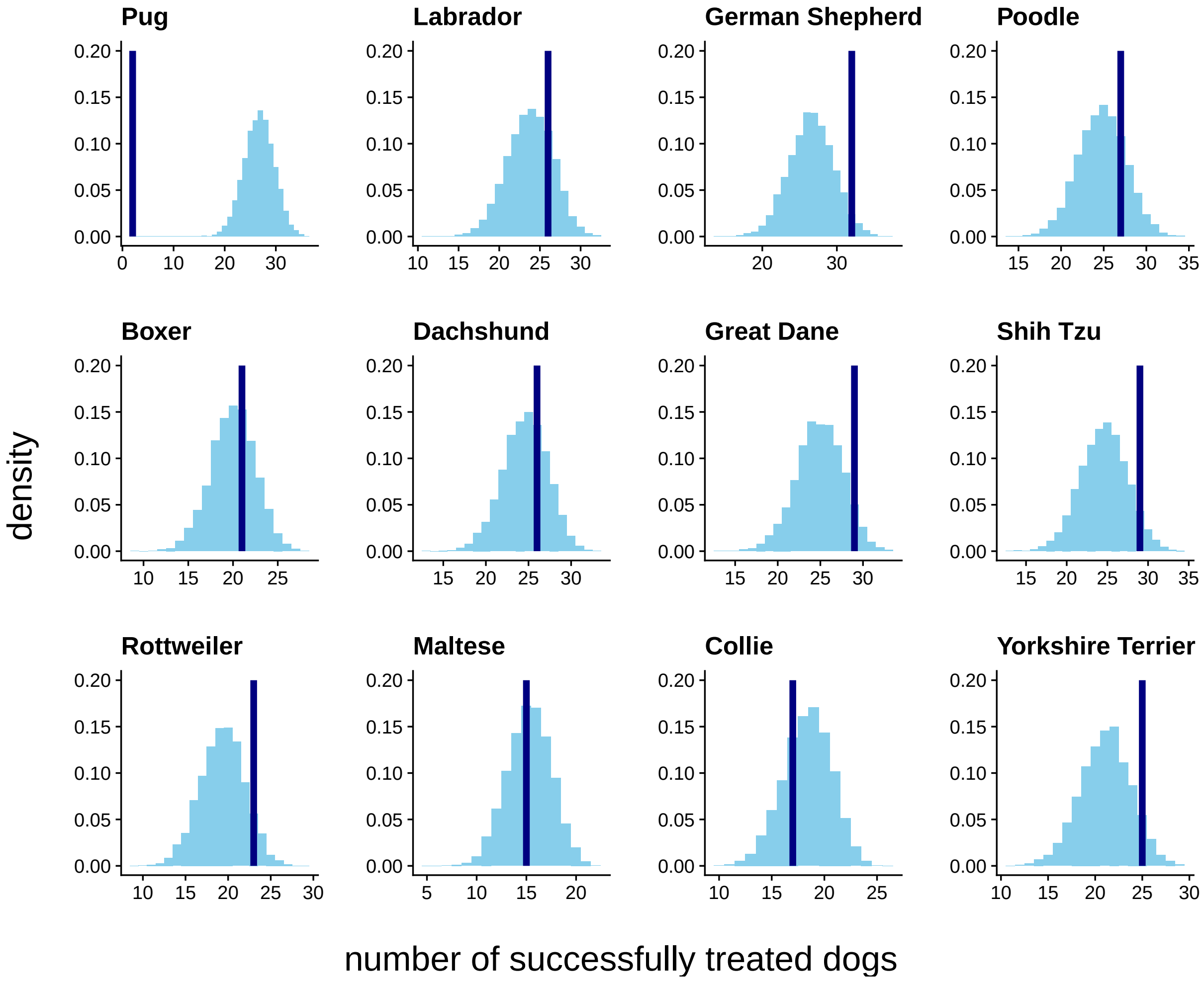


Figure 2.8: Posterior predictive check of the total number of successfully treated dogs under the model's assumptions (histograms) and observed total number of successfully treated dogs (vertical lines).

The plots show that for all recorded dog breeds the observed total number of successfully treated dogs is well within the range of values expected by model, with one exception: pugs. We observe only 3 successfully treated pugs in the dataset, but the model expects at least 15. How to deal with this discrepancy depends on the circumstances: A check of the data might reveal data entry errors for pugs (which should prompt another check for all data points), or we might simply deem that the observed discrepancy is irrelevant for our research goals (because we don't care about the treatment efficacy in pugs). However, it might also prompt us to revise the model. This is especially the case if we also want to estimate the effect of the insulin product in pugs. Maybe there is a biological difference between pugs and the other dog breeds that could explain why the insulin does not seem to be effective in dogs, for example because of changes in the insulin receptor.

For this example, we simply allow the model some additional flexibility and include a binary indicator variable that is 1 for pugs and 0 for all other dog breeds. This gives the model the required flexibility to predict lower treatment success probabilities for pugs than for the other dog breeds. Our updated model therefore is:

$$
\begin{aligned}
\alpha &\sim \text{Normal}(0, 1)\\
\beta_1 &\sim \text{Normal}(0, 1)\\
\beta_2 &\sim \text{Normal}(0, 0.2)\\
\beta_3 &\sim \text{Normal}(0, 0.1)\\
\beta_4 &\sim \text{Normal}(0, 1)\\
\\
y_i &\sim \text{Bernoulli}(\text{logistic}(\alpha + \beta_1 \times \text{height} + \beta_2 \times \text{weight} + \beta_3 \times \text{age} + \beta_4 \times \text{pug}))
\end{aligned}
$$

Figure 2.9 shows the posterior predictive check for the updated model. When including the pug indicator variable, the model is able to correctly capture the number of successfully treated pugs. Note how the improved model also better captures some other breeds, e.g. German Shepherd, Great Dane and Shih Tzu. Whether including a binary indicator variable for pugs is reasonable depends on our domain knowledge: If we really expect a biological reason why pugs are not responding to the insulin treatment it might make sense, otherwise it would probability be better to allow the treatment effect to vary for all dog breeds.

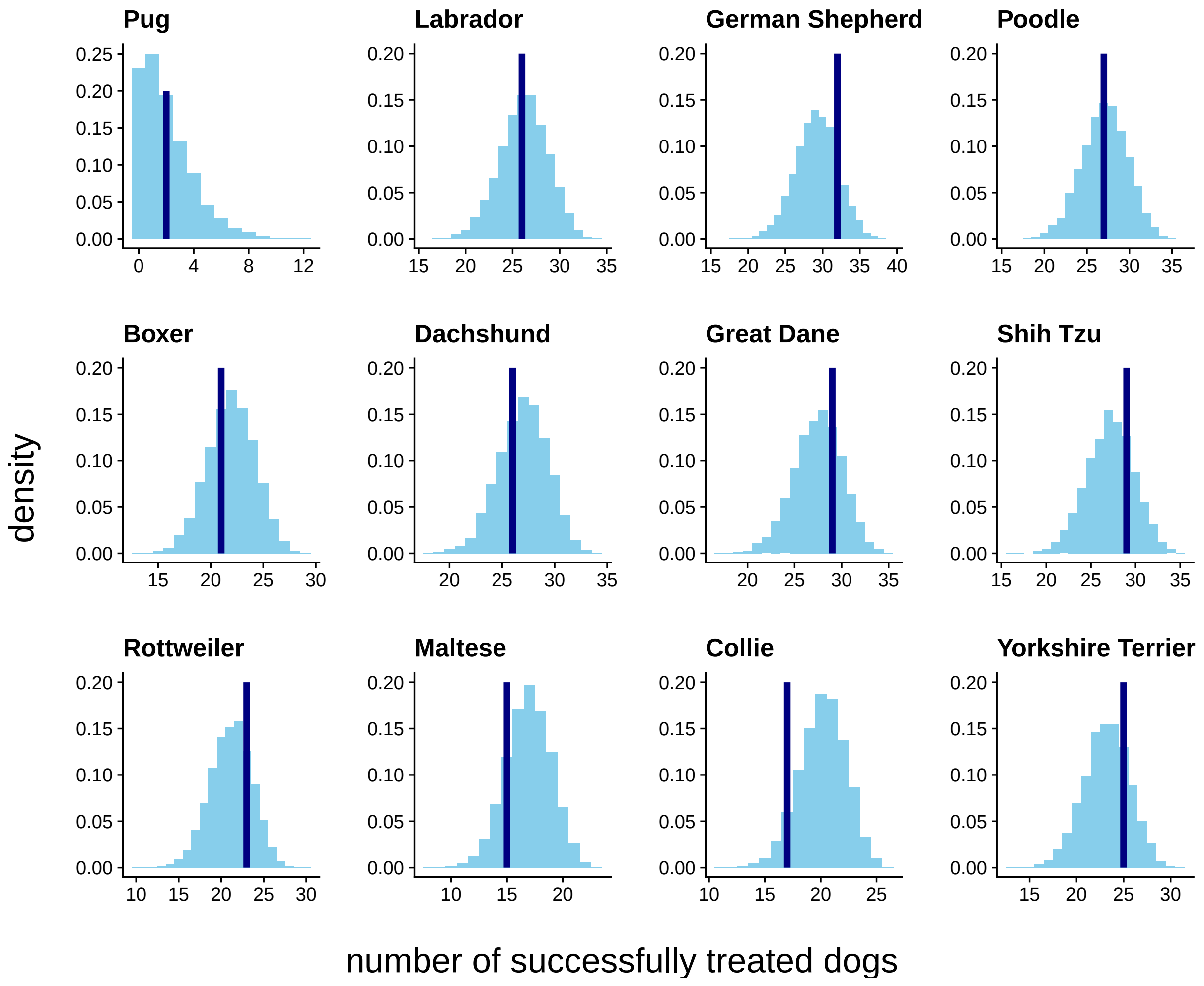


Figure 2.9: Posterior predictive check of the total number of successfully treated dogs under the model's assumptions (histograms) and observed total number of successfully treated dogs (vertical lines). The model including a binary indicator variable for pug shows no discrepancy between the expected and observed number of successfully treated pugs.

2.5.8 Pareto-smoothed importance sampling

After we have come up with a model that fits the observed data sufficiently, does not show any sampling issues and retrieves correct parameter values based on simulated data, the next step in Bayesian workflow is model evaluation. At this stage, the goal is to evaluate how well the model does perform according to metrics relevant to the application. For example in a prediction task, one might be interested in the proportion of correctly predicted data points, or in a regression task, a measure of accuracy like mean-squared error might be of interest. Metrics like classification accuracy and mean-squared error are closely related to model comparison: If a newly developed model is to replace a previous one, it is important to ensure that the new model actually outperforms the current model.

A straight-forward way to evaluate and compare models is to gather additional data:

If the models are applied to the new data, the obtained performance metrics are a good estimate of the model performance on future data. However, simply collecting more data is often not feasible, for example because it is too expensive. One might be inclined to test the models on the data they were fit on. Unfortunately, the obtained performance metrics are a poor estimate of the model performance on future data. This is because the observed data might not be representative enough, and future data has features that deviate from the current data. This phenomenon is most pronounced for very flexible models, e.g. models that have a lot of parameters to adapt to the observed data. In this case, a model might have stellar performance on the observed data, but poor performance on future data. Models like this are said to not *generalize*, and one term used to describe this issue is called *overfitting*.

One trick to alleviate this issue is called leave-one-out cross-validation (LOO-CV). In LOO-CV, a single data point is held out and the model is fit on all remaining data. The evaluation is then done on the held out data point, and this procedure is repeated until every data point is being held out once. As LOO-CV requires fitting the model multiple times (i.e. once for each data point), alternative approaches like k-fold cross-validation are sometimes used. In k-fold cross-validation, the data set is split into k groups (e.g. 10), and instead of one data point, one group is held out. This has the advantage that it requires less refitting of the model, but has the drawback that the model is also fit on fewer data.

In Bayesian statistics, the cost of refitting the model is often particularly high, because the methods used to draw samples from the posterior distribution are more computationally expensive. Fitting large models with many parameters might take hours, days, or even weeks. Therefore, computational methods have been developed that approximate the leave-one-out posterior based on the posterior obtained from the full data set. At its core, most of these methods are based on importance sampling, which is a method that can generate samples from one probability distribution (here: the leave-one-out posterior) based on samples obtained from a distribution that approximates the target distribution (here: the full posterior). The most important requirements for importance sampling to work are that a.) both distributions are valid on the same range of values (i.e. it is not possible to approximate a distribution that is defined for values between 0 and 1 with a distribution that is defined for all real values) and b.) that the density can be computed for arbitrary points.

Consider the following example that shows how importance sampling works: Suppose you have a computer program that is able to generate samples from a standard uniform distribution, but you need to generate samples from a Beta(2, 2) distribution instead. The most obvious solution is to use a sampling algorithm that specifically generates samples from a Beta distribution, but imagine these algorithms would not exist. One could then use the samples from the approximating distribution (in this case the standard uniform distribution) and use importance sampling to generate samples from the target distribution, which is done by weighted resampling. Intuitively speaking, sampling from the target distribution is achieved by calculating the density ratio between the target and approximating distribution. Samples for which the target distribution has a higher density than the approximating distribution are sampled more frequently (with a weight determined by the ratio), and regions were the target distribution has less density than the approximating distribution are sampled less often. As a concrete example, consider the samples obtained from the standard uniform distribution are 0.1, 0.5 and 0.9. The density of the uniform distribution $f(x)$ is 0.1 for all values, and the density of the Beta(2, 2) distribution $g(x)$ at locations 0.1, 0.5 and 0.9 is 0.54, 1.5 and 0.54, respectively. Therefore, the ratio of the densities $g(x)/f(x)$ are 5.4, 15, 5.4. Many programming languages have built-in functions for weighted resampling, e.g. in R, weighting resampling of 100 weighted draws can be achieved by `sample(c(0.1, 0.5, 0.9), 100, prob = c(5.4, 15, 5.4))`. Of course, resampling based on 3 data points is not useful. Figure 2.5.8 shows samples of a Beta(2, 2) distribution generated by resampling a uniform distribution based on 100 data points. The obtained histogram already follows the true density quite closely.

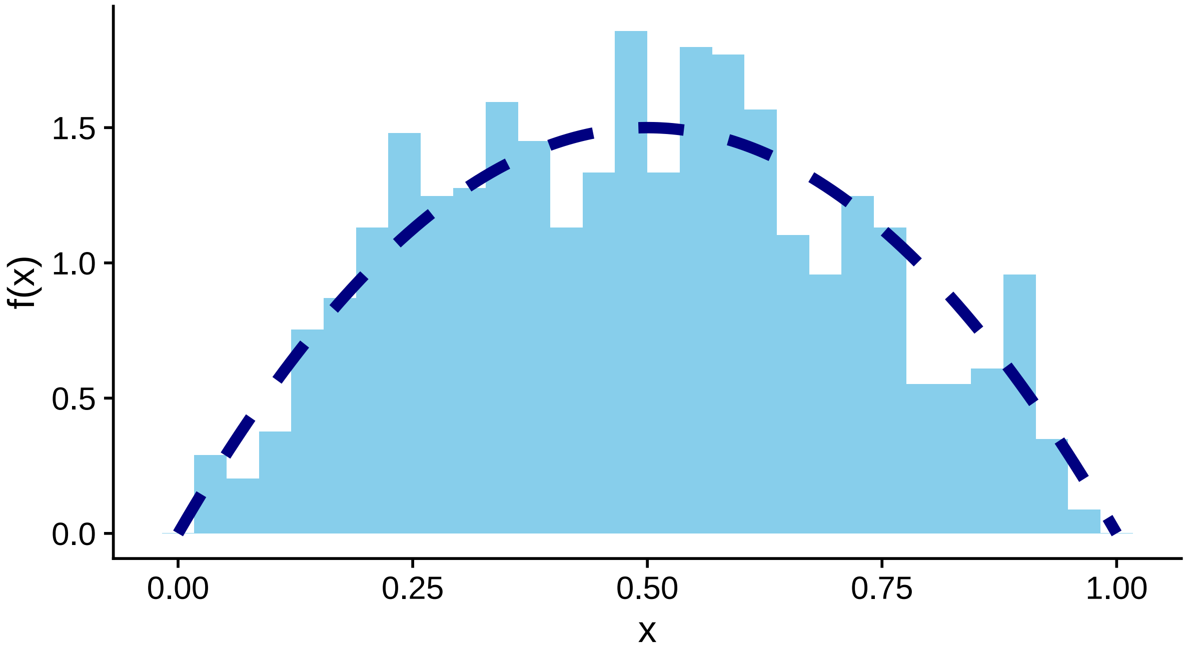


Figure 2.10:

The likelihood function of many Bayesian models can be written as the product of the individual point wise likelihoods, e.g. for a parameter vector θ, the posterior density (up to a constant) for many models can be written as $p(\theta)q_\theta(1)q_\theta(2)$, where $p(\theta)$ is the prior distribution and $q_\theta(1)$ is the likelihood for data point 1 given θ, $q_\theta(2)$ is the likelihood for data point 2 given θ and so on. Consider a model with 3 data points. The importance ratio for holding out data point 2 can therefore be computed as:

$$r = \frac{p(\theta)q_\theta(1)q_\theta(3)}{p(\theta)q_\theta(1)q_\theta(2)q_\theta(3)} = \frac{1}{q_\theta(2)} \tag{2.12}$$

Calculating importance ratios therefore only involves computing the likelihood function $p(y_i|\theta)$ at the observed data points. Importance sampling is useful because it allows evaluating models on held-out data points without refitting the model N times. However, like all approximations it might fail. This happens when the approximating distribution is not similar enough to the target distribution, so the importance ratios get really large or small and the resulting resampling becomes too noisy. Pareto-smoothed importance sampling (PSIS) improves this issue by fitting a Pareto distribution to the importance ratios and replacing the 20% largest importance ratios with importance ratios according to this Pareto distribution. This greatly stabilizes the importance sampling procedure because it is less sensitive to overly large importance ratios. It also has the advantage of providing an easy diagnostic metric: The parameter k of the Pareto distribution is useful to determine when Pareto-smoothed importance sampling might fail: Empirical data shows that PSIS is very accurate for k values below 0.7. For data points with k above 0.7, it is advised to refit the model and use the true leave-one-posterior instead.

We can use Pareto-smoothed importance sampling to compare our improved version of the insulin model which includes the pug predictor to the previous version. Table 2.1 shows the expected log-predictive density (elpd) for each model. The expected log-predictive density is a measure of model performance on future data. It denotes the expected probability density (or probability mass for discrete outcomes) that the model allocates to future data points, and we therefore generally prefer models with the highest elpd possible, because they are expected to predict future data best. The expected log-predictive density has the advantage over other measures of model performance that it not only takes the location of the estimates into account, but also the width of the predictive distribution (see Figure 2.11).

Table 2.1: Expected log-predictive density for the initial model and the model with the pug indicator variable. The elpd shows that the model with the additional predictor is expected to have a much better predictive performance on future data, as the difference in elpd between that model (-189.01) and the initial model (-232.63) is much higher than the standard errors of the estimates.

	elpd estimate	standard error
without pug predictor	-232.63	8.33
with pug predictor	-189.01	10.41

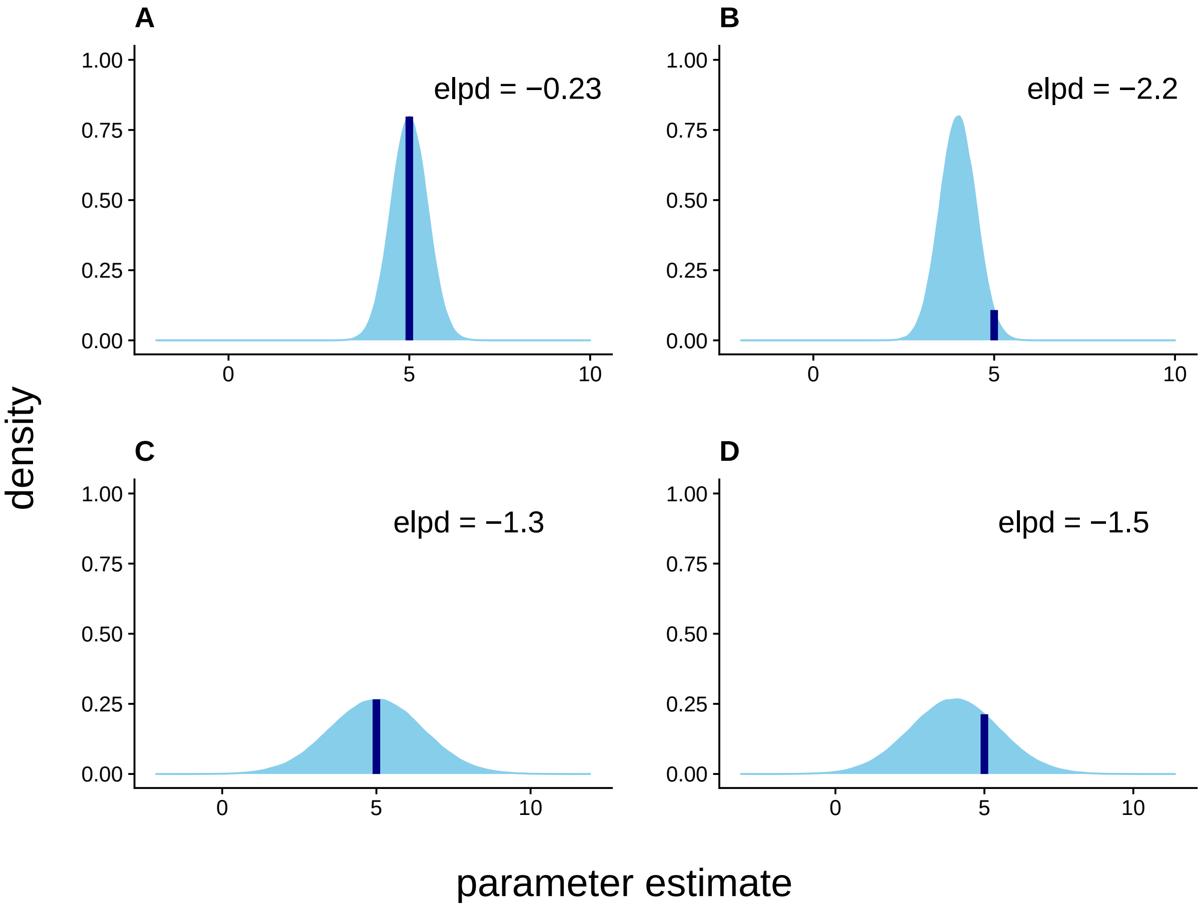


Figure 2.11: Expected log point-wise predictive density in four hypothetical scenarios. The vertical lines show an observed data point in the future (x = 5), the distributions show the posterior predictive distributions for that new data point. Expected log-predictive density combines the location and the uncertainty of the estimates: The highest elpd is achieved for models that produce estimates near the observed data point with little uncertainty (model A). Models that get either the location wrong or produce wider estimates (model B and model C) have lower elpd. The predictions of model D and model B have the same location, but because the posterior predictive distribution is wide enough, it has higher elpd than model B.

2.5.9 Model visualization

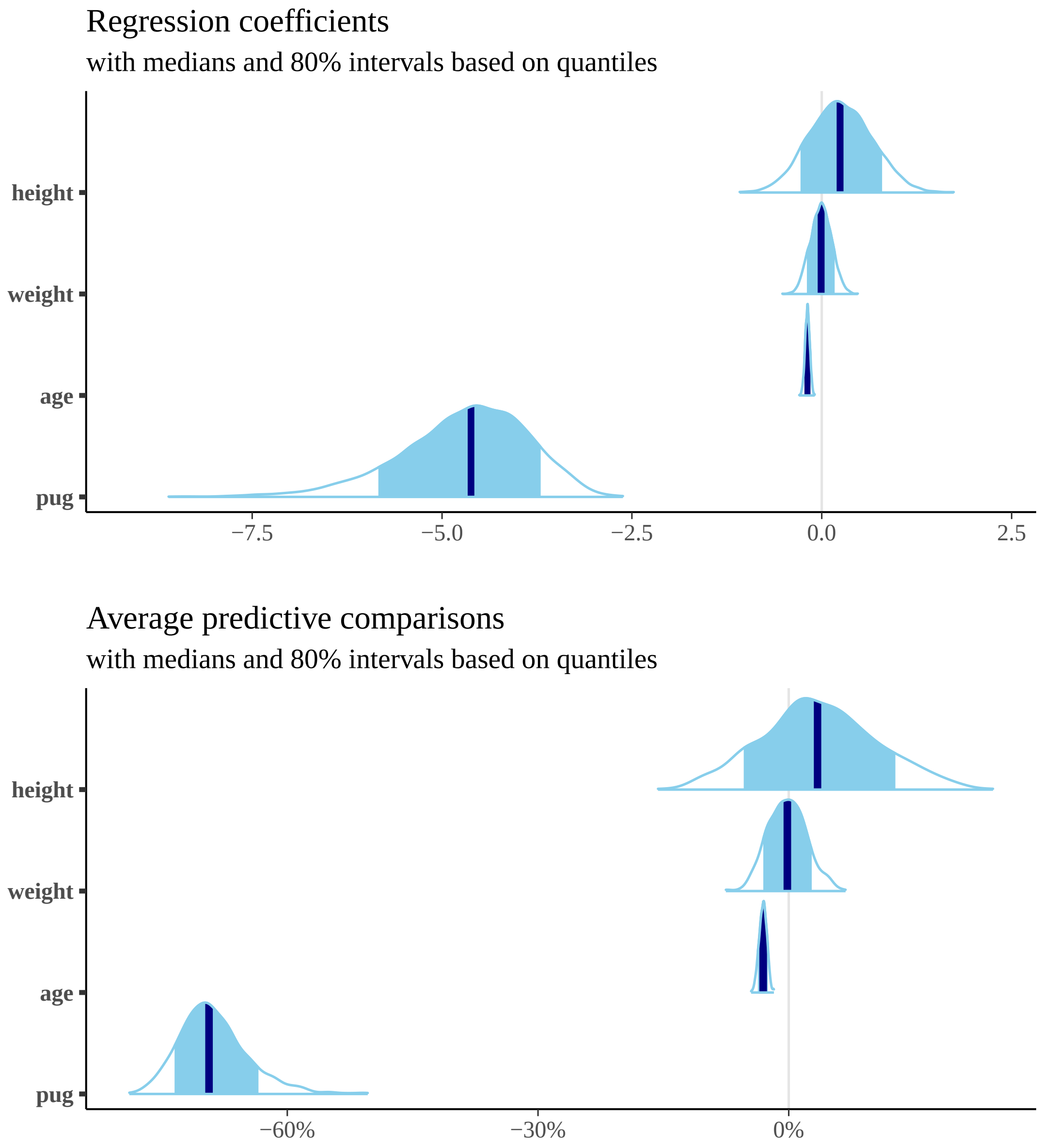
A common way to summarize the results of a regression model are tables of regression coefficients with uncertainty intervals. These tables have the advantage of displaying important information about the model in a compact manner, but the disadvantage that a.) it is not possible to compare regression coefficients in a large table at a glance, b.) regression coefficients alone are often difficult to interpret.

An immediate solution to a.) is to plot uncertainty intervals as horizontal or vertical bars. In this way, it is often much easier to compare different coefficients to each other. Possible solutions to the problem outlined in b.) are not as apparent: Consider a linear model like the previous regression of body height on weight: *body weight (in kg)* $= \beta_0 + \beta_{\text{height}} \times$ *body height (in cm).* In this model, the regression coefficient for body height β_{height} can be interpreted in two possible ways: In the first interpretation, the regression coefficient is seen as the associated change in outcome for a (hypothetical) one-unit change in the input. For the body weight example, this means that the coefficient β_{height} is the expected increase in body weight (in kg), if body height would increase by 1 cm. In the second interpretation, the regression coefficient is seen as the average difference between two groups that differ by one unit in their predictors. For the body weight example, this means that β_{height} is the average body weight difference (in kg) between two groups of people, where one group is on average 1 cm taller than the other. Which of the two interpretations is more useful is context specific: For models in which a change in one of the inputs is impossible (like in the body height example) the difference-between-groups interpretation might make more sense, for models in which a more causal interpretation is desired, the one-unit-change interpretation might be preferred.

These two ways of interpreting regression coefficients also work for models with multiple inputs. Consider adding a binary variable for sex to the body weight model: *body weight (in kg)* $= \beta_0 + \beta_{\text{height}} \times$ *body height (in cm)* $+ \beta_{\text{sex}} \times$ *sex (1 female / 0 male).* The regression coefficient for β_{height} can still be interpreted as the associated change in weight for an increase in height by 1 cm, and it can also be interpreted as the average difference between groups that differ in their average height by 1 cm. However, both interpretations now require the additional assumption that all other inputs are held constant, e.g. β_{height} is the average difference between two groups that differ in their average height by 1 cm, with the additional requirement that these groups are either all male or all female.

Regression coefficients for models with non-linear link functions like logistic regression models can also be difficult to interpret: In logistic regression models, the regression coefficients are on the log-odds scale, which means that a one unit change in one of the inputs is associated with an increase in log-odds of observing the outcome denoted $y_i = 1$ by the value of the regression coefficient. Log-odds are defined as $\log(p/1-p)$. Defining the model in terms of log odds has the advantage that the individual predictors contribute additively to the outcome, but the disadvantage that it is difficult to reason about odds, as we are more used to probabilities in our every day life.

A method that can make the interpretation of regression coefficients easier for models with interactions, non-linear predictors or regression coefficients in unwieldy units like in logistic regression is called *Average Predictive Comparisons* (Gelman and Pardoe, 2007). Predictive Comparisons denote the expected change in the output by a one unit change in one of the inputs. This in general depends on the beginning and end points of the hypothesized change, the values of the other inputs and the parameters of the model. *Average* Predictive Comparisons average over these changes to obtain the expected change in the output by a one unit change in one of the inputs. For linear models without interaction terms average predictive comparisons are equal to regression coefficients. Figure 2.12 shows regression coefficients (on the scale of log-odds) and average predictive comparisons (on the probability scale).


Regression coefficients

with medians and 80% intervals based on quantiles

Average predictive comparisons

with medians and 80% intervals based on quantiles

Figure 2.12: Density plots showing the regression coefficients for each parameter (top) and average predictive comparisons for each input (bottom). For logistic regression models, the regression coefficients are on the scale of log-odds, i.e. for pugs, the log-odds of treatment success is decreased by about 3 compared to the other dog breeds. If we use the model parameters and data to reframe this in terms of average predictive comparisons, pugs are about 70% less likely to respond to the treatment, compared to the other dog breeds. It is usually much easier to reason about effects on the probability scale than on the log-odds scale.

2.5.10 Summary

This example showed a brief overview of a Bayesian workflow. A full description of the Bayesian workflow can be found in Betancourt (2020); Gabry et al. (2019); Gelman et al. (2020). It is important to realize that the "Bayesian workflow" is not set in stone, parts of it can be left out and the focus can be changed according to the needs of the specific analysis.

The Bayesian workflow can be seen as a shift in perspective: Instead of applying a statistical method to data like a recipe from a cookbook, we focus on repeated cycles of model fitting and model evaluation, being guided by principles that follow from our research goals. We lose some rigor in the sense that we let our decisions be informed by the observed data, but in practice, this drawback is easily compensated by the fact that we end up with models that do not exhibit strong misfits to the data.

However, following this principle also has some downsides: It cannot be applied blindly and requires statistical knowledge and might therefore be not as easily applicable as standard methods. This downside can also be an advantage however, because practitioners who use statistics as a research tool can obtain valuable feedback if the applied methods are valid to answer the research question. After some initial training, Bayesian methods can actually be more accessible, because a.) following the workflow does not require the same background in probability theory as designing a hypothesis test from the grounds up does, and b.) after some time, the pieces just "all fit together" and applying a new procedure is as simple as reordering parts that are already known. This reduces the black-box character of statistics and while certainly not a panacea, can greatly help to obtain more reliable insights from data.

Chapter 3

Contributed Articles

3.1 HAMdetector: A Bayesian regression model that integrates information to detect HLA-associated mutations

This section is based on the following publication:

Daniel Habermann, Hadi Kharimzadeh, Andreas Walker, Yang Li, Rongge Yang, Rolf Kaiser, Zabrina L. Brumme, Jörg Timm, Michael Roggendorf, Daniel Hoffmann (2022). **HAMdetector: A Bayesian regression model that integrates information to detect HLA-associated mutations.**
`https://doi.org/10.1093/bioinformatics/btac134`

Bioinformatics, 2022, 1–9
https://doi.org/10.1093/bioinformatics/btac134
Advance Access Publication Date: 3 March 2022
Original Paper

Sequence analysis

HAMdetector: a Bayesian regression model that integrates information to detect HLA-associated mutations

Daniel Habermann [1,*], Hadi Kharimzadeh[2], Andreas Walker[3], Yang Li[4], Rongge Yang[4], Rolf Kaiser[5], Zabrina L. Brumme[6,7], Jörg Timm[3], Michael Roggendorf[8] and Daniel Hoffmann [1,9,10,*]

[1]Bioinformatics and Computational Biophysics, Faculty of Biology, University of Duisburg-Essen, Essen 45117, Germany, [2]Division of Clinical Pharmacology, University Hospital, LMU Munich, Munich, Germany, [3]Institute of Virology, Medical Faculty, University Hospital Düsseldorf, Heinrich-Heine-Universität, Düsseldorf 40225, Germany, [4]AIDS and HIV Research Group, State Key Laboratory of Virology, Wuhan Institute of Virology, Chinese Academy of Science, Wuhan, China, [5]Institute of Virology, University of Cologne, Faculty of Medicine and University Hospital of Cologne, Cologne 50935, Germany, [6]Faculty of Health Sciences, Simon Fraser University, Burnaby, Canada, [7]British Columbia Centre for Excellence in HIV/AIDS, Vancouver, Canada, [8]Institute of Virology, School of Medicine, Technical University of Munich/Helmholtz Zentrum München, Munich, Germany, [9]Center of Medical Biotechnology, University of Duisburg-Essen, Essen, Germany and [10]Center for Computational Sciences and Simulation, University of Duisburg-Essen, Essen, Germany

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on July 22, 2021; revised on November 21, 2021; editorial decision on February 11, 2022; accepted on February 28, 2022

Abstract

Motivation: A key process in anti-viral adaptive immunity is that the human leukocyte antigen (HLA) system presents epitopes as major histocompatibility complex I (MHC I) protein–peptide complexes on cell surfaces and in this way alerts CD8+ cytotoxic T-lymphocytes (CTLs). This pathway exerts strong selection pressure on viruses, favoring viral mutants that escape recognition by the HLA/CTL system. Naturally, such immune escape mutations often emerge in highly variable viruses, e.g. HIV or HBV, as HLA-associated mutations (HAMs), specific to the hosts MHC I proteins. The reliable identification of HAMs is not only important for understanding viral genomes and their evolution, but it also impacts the development of broadly effective anti-viral treatments and vaccines against variable viruses. By their very nature, HAMs are amenable to detection by statistical methods in paired sequence/HLA data. However, HLA alleles are very polymorphic in the human host population which makes the available data relatively sparse and noisy. Under these circumstances, one way to optimize HAM detection is to integrate all relevant information in a coherent model. Bayesian inference offers a principled approach to achieve this.

Results: We present a new Bayesian regression model for the detection of HAMs that integrates a sparsity-inducing prior, epitope predictions and phylogenetic bias assessment, and that yields easily interpretable quantitative information on HAM candidates. The model predicts experimentally confirmed HAMs as having high posterior probabilities, and it performs well in comparison to state-of-the-art models for several datasets from individuals infected with HBV, HDV and HIV.

Availability and implementation: The source code of this software is available at https://github.com/HAMdetector/Escape.jl under a permissive MIT license. The data underlying this article were provided by permission. Data will be shared on request to the corresponding author with permission of the respective co-authors.

Contact: daniel.habermann@uni-due.de or daniel.hoffmann@uni-due.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

1.1 The human leukocyte antigen system

The human immune system recognizes viral infections through two pathways: The innate and adaptive immune response. T-cell, or 'cellular', immunity, which represents one major arm of the adaptive immune system, is modulated by human leukocyte antigen (HLA) molecules (Germain, 1994): Briefly, proteins that are synthesized within the cell—which will include viral proteins if the cell is infected—are degraded in proteasomes to peptides (Goldberg *et al.*, 2002). Some of these peptides are presented as epitopes on the cell surface by HLA class I molecules. These viral peptide-HLA complexes can then be recognized by circulating CD8$^+$ cytotoxic T-lymphocytes (CTLS) through their T-cell receptor (Murata *et al.*, 2007). Following this recognition, the CTL can eliminate the infected cell (Harty *et al.*, 2000).

HLA class I molecules are encoded at three loci, HLA-A, -B and -C and these genes are very polymorphic with more than 20 000 known alleles in humans (Robinson *et al.*, 2015). HLA molecules vary drastically in their affinities to given epitopes so that cells from different individuals, in general, present different peptides on the cell surface. In other words, the HLA class I alleles expressed by a given individual will determine their CTL response to a given viral pathogen.

1.2 Immune escape is reproducible based on host HLA allele expressed

Virus variants arise continuously through mutation. Because the HLA system modulates CTL responses through viral epitope presentation, it exerts strong selection pressure toward virus variants that escape CTL recognition (Borrow *et al.*, 1997). Such variants could, for example, carry mutations that reduce binding of viral epitopes to HLA, or that reduce recognition of the epitope/HLA complex by the CTL's T-cell receptor, or that alter peptide processing so that epitopes are no longer presented on the infected cell surface (Yewdell *et al.*, 2002). The latter type of mutation can occur within (Yokomaku *et al.*, 2004) or outside (Draenert *et al.*, 2004) CTL epitopes.

Immune escape is a major driver of viral evolution, particularly for highly variable viruses such as HIV or HBV (Alizon *et al.*, 2011; Allen *et al.*, 2005; Lumley *et al.*, 2018; Rousseau *et al.*, 2008). Whether and how quickly a given escape mutation is selected in a host depends on a number of factors including the viral genomic background, the magnitude of the reduction in viral replication caused by changes in the viral proteins, the selection of compensatory mutations that recover fitness, and the strength of immune response targeting the presented epitope (Kløverpris *et al.*, 2015). Despite the complexity of these factors, the mutational pathways of immune escape in certain viruses such as HIV are nevertheless broadly reproducible, and thus predictable, based on the HLA alleles expressed by the host. For example, about 75% of people living with HIV who carry the HLA-B*57 allele, will select a T242N substitution in the HIV structural protein Gag in the first weeks to months of infection (Brumme *et al.*, 2008b; Leslie *et al.*, 2004).

In addition to driving viral evolution at the individual level, HLA pressures also drive viral evolution in human populations, as circulating viruses adapt to HLA alleles commonly expressed in that population (Kawashima *et al.*, 2009). Upon transmission to a new host with different HLA alleles, HLA escape mutations may revert, particularly if they are associated with a reduction in viral replication capacity (Matthews *et al.*, 2008), but they can also persist, leading to their population-level accumulation (Kawashima *et al.*, 2009).

Methods to accurately and comprehensively identify HLA-associated immune escape mutations in HIV and other viruses are therefore critical for the study of viral evolution and immune escape. An improved understanding of immune escape can aid in the development of treatments and vaccines that rely on effective immune responses.

1.3 Identifying HLA escape mutations

There are several experimental methods available to study HLA escape (Altman *et al.*, 1996; Brunner *et al.*, 1968; Czerkinsky *et al.*, 1983; Lamoreaux *et al.*, 2006). However, these methods are relatively slow and costly, especially for screening purposes. Theoretically, an option to identify escape mutation could be the use of epitope prediction tools (Mei *et al.*, 2020). At their core, these tools identify epitopes as peptides with high predicted affinities to HLA molecules. One could envisage applying such tools to combinatorially mutated epitopes to identify substitutions that reduce predicted affinities significantly and therefore would be good candidates for escape mutations. However, these tools can be rather insensitive to escape mutations (Acevedo-Sâenz *et al.*, 2015), which is not unexpected because they have been developed to recognize epitopes as a whole. A more promising approach that makes efficient use of frequently available data is to combine viral genome sequencing, host HLA determination, computational identification by statistical association analysis and targeted experimental validation (Carlson *et al.*, 2012).

As the selection pressure exerted by cytotoxic T cells depends on successful recognition of viral peptides bound to HLA molecules on the infected cell surface, escape mutations are HLA allele specific and can therefore be detected as HLA allele dependent amino acid substitutions, or 'footprints,' in sequence alignments of viral proteins (Moore, 2002). Amino acid substitutions enriched in viral sequences from hosts with a specific HLA allele are termed HLA-associated mutations (HAM).

One way of quantifying this enrichment is Fisher's exact test (Fisher, 1922): For a given substitution S_i at alignment position i and HLA allele H, a 2-by-2 contingency table is constructed containing the absolute counts of the number of sequences in the four possible categories (S_i, H), $(S_i, \neg H)$, $(\neg S_i, H)$ and $(\neg S_i, \neg H)$, where $\neg S_i$ denotes any substitution except S_i, and $\neg H$ denotes any HLA allele except H.

Fisher's exact test is a conventional null hypothesis significance test (NHST) that generates P-values. In this case, the null hypothesis is that HLA allele H and substitution S_i are independent, and the P-value is the probability of observing a deviation from independence that is at least as extreme as in the data at hand under the assumption that the null hypothesis is true.

Fisher's exact test has the advantage of being fast and easy to apply (Budeus *et al.*, 2016), but it also has several disadvantages (Carlson *et al.*, 2008). The most striking one is that viral sequences share a common phylogenetic history, and, therefore, treating sequences as independent and identically distributed samples may under- or overestimate effect sizes. In the context of hypothesis testing, this leads to increased false-positive and false-negative rates (Osborne *et al.*, 2002; Scariano *et al.*, 1987).

Another issue with Fisher's exact test is the genomic proximity of human HLA class I loci (Francke *et al.*, 1977) leading to linkage disequilibrium—inheritance of HLA alleles can be correlated. Therefore, spurious HAMs can occur if associations of substitutions with individual HLA alleles are tested: if HLA allele H_1 is associated with an amino acid substitution R because of immune escape, but H_1 is in linkage disequilibrium with allele H_2, then this leads to an association of R and H_2, even without being an escape mutation from H_2.

Carlson *et al.* (2008) developed the Phylogenetic Dependency Network, a method that accounts for several of the aforementioned problems, in particular phylogenetic bias and HLA linkage disequilibrium. However, it is based on null hypothesis significance testing.

1.4 Issues with *P*-values for screening

There are fundamental statistical issues with P-values as a screening tool (Amrhein et al., 2018): with small effect sizes and high variance between measurements, as is often the case with biological data, statistically significant results can be misleading, can have the wrong direction (type S error), or can greatly overestimate an effect (type M error) (Gelman *et al.*, 2014). Such problems are more and more appreciated in the context of the current 'replication crisis'—in the life sciences scientific claims with seemingly strong statistical support often fail to replicate (Baker, 2016; Begley *et al.*, 2012; Ioannidis, 2005).

These problems are exacerbated if *P*-values are used for screening purposes (multiple testing problem). The probability of obtaining a statistically significant result increases with each additional test, even in absence of any real effect. When using *P*-values as a filter, it is therefore likely to obtain significant effects that are in fact not real. A common strategy to mitigate this problem is to control the false discovery rate (Benjamini *et al.*, 1995). The downside of such adjustment procedures is that only the very largest effects remain if large datasets are screened.

Instead of performing many hypothesis tests and trying to adjust for them, we prefer to fit a single, multilevel model that contains all comparisons of interest. Multilevel models can make the problem of multiple comparisons disappear entirely and yield more valid estimates (Gelman *et al.*, 2012).

2 Materials and methods

Our general approach for HAMdetector is to fit a Bayesian regression model that captures relationships between host HLA alleles and substitutions in viral proteomes.

This Bayesian approach is advantageous because it allows use of: (i) prior information (e.g. knowledge of effect magnitudes), (ii) relevant additional information (phylogeny, epitope information), (iii) a problem-specific structure and (iv) partial pooling (Gelman, 2010).

2.1 Model backbone

We chose a logistic regression model as backbone because it is easily extensible, and because coefficients can be interpreted in the familiar way as summands on the log-odds scale. This is the core of HAMdetector, which models the strength of association between substitutions in viral sequences and host HLA alleles.

$$y_{ik} \sim \text{Bernoulli}(\theta_{ik}) \tag{1}$$

$$\theta_{ik} = \text{logistic}(\beta_{0_k} + \sum_{j=1}^{D} X_{ij}\beta_{jk}), \tag{2}$$

where y_{ik} is the binary encoded observation of substitution k in viral sequence i (each observed amino acid state k contributes a separate column to y_{ik}); θ_{ik} is the estimated probability that we observe substitution k in sequence i; β_{0k} is an intercept for substitution k, corresponding to the overall log-odds for substitution k; X_{ij} is 1 if sequence i comes from host individual with HLA allele j and 0 otherwise; β_{jk} is the HLA regression coefficient of HLA allele j for substitution k; D is the number of HLA alleles in the dataset; the logistic inverse link function transforms the linear model in parentheses to the probability scale of θ_{ik}.

The main parameters of interest for HAMdetector are the regression coefficients β_{jk}, as they quantify the strength of association between the occurrence of substitution k and each of the observed HLA alleles. The β_{jk} are on the log-odds scale, i.e. if we go from viral sequences from hosts without HLA allele j to those from hosts with j, the log-odds $\log(p_k/(1-p_k))$ of observing substitution k increase by addition of β_{jk}.

Reasoning about coefficients on the log-odds scale can sometimes be unintuitive. A useful approximation to interpret logistic regression coefficients on the probability scale is the so-called divide-by-4 rule, which means that a regression coefficient of 2 corresponds to an expected increase on the probability scale of up to 2/4 = 50%.

2.2 Inclusion of additional information

On top of the paired data of viral sequences and host HLA alleles modeled by the backbone (Eq. 1), we extend the model to include further information of relevance to improve HAM detection, namely phylogenetic information and predictions of epitope peptide processing and major histocompatibility complex I (MHC I) affinity, as described in the following.

2.2.1 Phylogeny

Viral strains have a common phylogenetic history. Thus substitutions are not independently and identically distributed, and therefore violate a common assumption of standard statistical methods. In fact, Bhattacharya *et al.* (2007) demonstrated the importance of correcting for the phylogenetic structure in identifying HLA associations.

A popular approach in phylogeny-aware regression of binary variables is to estimate an additional multivariate normally distributed intercept, where the covariance matrix is based on the branch lengths of a given phylogenetic tree (Ives *et al.*, 2010, 2014). This approach turned out to be too computationally expensive in our model, hence we chose a strategy similar to the one in Carlson *et al.* (2008):

Consider a phylogenetic tree Ψ obtained from standard maximum likelihood methods for a given multiple sequence alignment. We are interested in estimating $P(y_{ik} = 1|\Psi)$, that is, the probability of observing the substitution k in sequence i based on the underlying phylogenetic model. A quantity that can be readily computed using phylogenetic software like RAxML-NG (Kozlov *et al.*, 2019) is $P(\Psi|y_{ik} = 1)$. For this, we keep the tree topology fixed, annotate the tree with the binary observations y_{ik} at its leaves and optimize the branch lengths. $P(\Psi|y_{ik} = 1)$ is then the likelihood of the annotated phylogenetic tree. Similarly, we can also compute $P(\Psi|y_{ik} = 0)$ by flipping the annotation of sequence i from 1 to 0 (keeping all other observations). With $P(\Psi|y_{ik} = 1)$ and $P(\Psi|y_{ik} = 0)$ known and the relative frequencies of 0 and 1 as priors, we can estimate $P(y_{ik} = 1|\Psi)$ by applying Bayes' theorem. The estimated probabilities based on phylogeny are then included in the model as additional intercepts (second term of logistic argument):

$$\begin{aligned} y_{ik} &\sim \text{Bernoulli}(\theta_{ik}) \\ \theta_{ik} &= \text{logistic}(\beta_{0_k} + \gamma \text{logit}\Big(P(y_{ik} = 1|\Psi)\Big) \\ &\quad + \sum_{k=1}^{D} X_{ik}\beta_{jk}) \end{aligned} \tag{3}$$

The logit transform is used because it cancels out with the logistic inverse link function. The phylogeny term acts as a baseline in absence of any HLA effects. As this baseline itself is not certain but subject to errors of the phylogenetic probabilities $P(y_{ik} = 1|\Psi)$, we introduce an additional parameter γ.

2.2.2 Inclusion of CTL epitope predictions

As outlined earlier, escape mutations often appear as HAMs. Given the underlying mechanism, it is not surprising that escape mutations are enriched in CTL epitopes, i.e. in those viral peptides presented by MHC I to TCRs (Bronke *et al.*, 2013). This suggests that knowledge of epitope regions can be used to boost HAM detection. Fortunately, availability of large experimental datasets (Vita *et al.*, 2019) has enabled the development of computational tools that predict with good accuracy the binding of peptides to MHC I molecules encoded by various HLA alleles (Mei *et al.*, 2020).

Not only mutations in CTL epitopes can lead to failure to present epitopes to T cell receptors, but also mutations at epitope-flanking positions that interfere with pre-processing of peptides, notably proteasomal cleavage of viral proteins (Le Gall *et al.*, 2007; Milicic *et al.*, 2005).

In HAMdetector, we use MHCflurry 2.0 (O'Donnell *et al.*, 2020) to predict epitopes that are properly processed and presented by MHC I. For this, we create an input matrix of dimensions $R \times D$, where R is the number of evaluated substitutions and D is the number of observed HLA alleles in the dataset. The elements of this matrix are binary encoded and contain a 1 if that position is predicted to be in an epitope, and 0 otherwise. Given an amino acid sequence, MHCflurry provides a list of possible epitopes (9–13 mers) and HLA allele pairs and calculates a rank based on comparisons with random pairs of epitopes and HLA alleles. For the binarization, we use the rank threshold of 2% suggested by MHCflurry.

We use epitope prediction as information about the expected degree of sparsity, i.e. if we know that there is an epitope restricted by a given HLA allele at that location, we expect that this HLA allele is more likely to be associated with substitutions at that position than the other HLA alleles. This idea is implemented by increasing the scale of the local shrinkage parameters λ_{jk} depending on epitope information:

$$\begin{aligned} \lambda_{jk} &\sim \text{Cauchy}^{+}(0, \sigma_j \exp(Z_{jk}\beta_{\text{epi}})) \\ \beta_{\text{epi}} &\sim \text{Normal}^{+}(1, 2), \end{aligned} \tag{4}$$

where Z_{jk} is 1 if HLA allele j is predicted to restrict the alignment position corresponding to substitution k, and 0 otherwise. The parameter β_{epi} governs the increase in scale of the corresponding local shrinkage parameters. The larger the estimated values of β_{epi} are, the more likely it is to see non-zero regression coefficients for these HLA alleles.

2.2.3 Sparsity-inducing priors

Sparsity-promoting priors (Piironen *et al.*, 2017b) can drastically improve predictive performance, because the model is better able to differentiate between signal and noise. These priors convey the a priori expectation that most coefficients in a regression model are close to 0, i.e. that non-zero coefficients are sparse. This assumption is likely correct for HAMs: the dominating mechanism that leads to HLA association of mutations is probably selection of mutations that mediate escape from MHC I presentation of epitopes; however, we know that these epitopes are sparse, i.e. the number of actual epitopes that are restricted by a given HLA allele is typically small compared to the number of all conceivable epitopes. Thus, for most pairs of HLA allele and substitution, the association is likely truly zero. Note that this reasoning does not preclude associations outside of epitopes as sometimes observed for compensatory mutations (Ruhl *et al.*, 2012) but just implies that these are more rare.

There is a range of sparsity-promoting priors with slightly different properties. They share the common structure of placing most probability mass very close to 0, with heavy tails to accommodate the non-zero coefficients. For our model, we use the so-called regularized horseshoe prior (Piironen *et al.*, 2017b), which is an improvement of the original horseshoe prior presented by Carvalho *et al.* (2010), in that it additionally allows some shrinkage for the non-zero coefficients. The original horseshoe prior is given by:

$$\begin{aligned} \beta_{jk} &\sim \text{Normal}(0, \tau^2\lambda_{jk}^2) \\ \lambda_{jk} &\sim \text{Cauchy}^{+}(0, 1) \\ \tau &\sim \text{Cauchy}^{+}(0, \tau_0), \end{aligned} \tag{5}$$

where β_{jk} are the regression coefficients; τ and λ_{jk} are the so-called global and local shrinkage parameters, respectively; Cauchy^{+} is the positively constrained Cauchy distribution; τ_0 is the overall degree of sparsity. Shrinkage of the non-zero coefficients in the regularized horseshoe prior is achieved by replacing λ_{jk}^2 with $\tilde{\lambda}_{jk}^2 = \frac{c_k^2\lambda_{jk}^2}{c_k^2+\tau^2\lambda_{jk}^2}$, where the additional parameter c governs the magnitude of shrinkage for the non-zero coefficients.

With Eq. 5, the global shrinkage parameter τ is typically very small and shrinks most of the regression coefficients close to 0, whereas the local shrinkage parameters λ_{jk} can occasionally be very large to allow some coefficients to escape that shrinkage.

The overall degree of sparsity τ_0 can be chosen based on the expected number of non-zero coefficients (Piironen *et al.*, 2017a).

The full model specification together with a prior justification is given in Supplementary Information.

2.3 Model implementation

A Julia (Bezanson *et al.*, 2017) package is available at https://github.com/HAMdetector/Escape.jl to run the model on custom data. Due to restrictions of dependencies (MHCflurry and RAxML-ng), HAMdetector is currently only available on Linux, but can be run on Windows using the Windows Subsystem for Linux (WSL2). All models were implemented in Stan 2.23 (Stan Development Team, 2021), a probabilistic programming language and Hamiltonian Monte Carlo sampler for efficient numerical computation of posterior distributions. The Stan code is available in two versions: One optimized for readability and one optimized for speed by utilizing Stan's multithreading and GPU capabilities.

2.4 Model diagnostics

2.4.1 Convergence diagnostics

We use the split-$\hat{R}$ convergence diagnostic to identify Markov chain convergence issues (Gelman *et al.*, 1992, 2013). We require a value of $\hat{R}$ below 1.1 for all model parameters. Additionally, we require that the effective sample size N_{eff} (Stan Development Team, 2021) is above 200 for all model parameters and that sampling occurs without any divergent transitions (Betancourt, 2018).

2.4.2 Posterior predictive checks

In posterior predictive checks, we simulate new data from the inferred posterior distribution and the likelihood, and we compare these simulated data with representative real data (Gabry *et al.*, 2019). A good model should predict data that are consistent with real data. This general idea was employed in two ways to test our models.

For a first posterior predictive check we used *calibration plots* (Supplementary Fig. S1): two binned quantities were plotted against each other, the observed relative frequencies of substitutions $f(y_{ik} = 1)$, and the predicted probabilities $P(y_{ik} = 1|\text{model})$. In such a plot, a well-calibrated model should yield points following the diagonal. Technically, all observations were first sorted by increasing estimated probability $P(y_{ik} = 1|\text{model})$ and grouped into n bins. For each bin, the fraction of observations with $y_{ik} = 1$ (observed event percentage) was then plotted against the midpoint of each bin. The cutpoints of the bins are indicated by error bars.

Second, we assessed the abilities of different models and methods to discover HAMs with *HAM enrichment plots*. These plots are based on the observation that CTL escape mutations are enriched in epitopes (Bronke *et al.*, 2013). Hence, the degree by which methods for HAM prediction recover this trend is a measure of model performance. To implement this measure, we first ranked all evaluated substitutions according to their respective credibility of being a HAM, computed as integral of the marginal posterior $P(\beta_{jk} > 0)$. For comparison with established methods, namely Fisher's exact test and Phylogenetic Dependency Network (Carlson *et al.*, 2008), ranked lists based on P-values were computed. Then we calculated for each rank r the accumulated number $N_e(r)$ of predictions of this rank or better ranks were located inside known epitopes. The higher the curve $N_e(r)$, the higher the enrichment of predicted HAMs in epitopes, see e.g. Figure 1.

2.4.3 Leave-one-out cross-validation

Another performance measure is the ability to generalize to unseen data. To examine this ability for the different model variants we performed leave-one-out cross-validation (LOOCV), using the efficient Pareto-smoothed LOOCV (Vehtari *et al.*, 2017).

From the LOOCV, we obtain the Expected Log-Predictive Density (ELPD) $\sum_{i=1}^{n} \log\left(\int p(y_i|\theta)p(\theta|y_{i-1})d\theta\right)$ for samples $i = 1, \ldots, n$, ith observation y_i, data y_{i-1} with the ith data point left out, and model parameters θ. Thus, the ELPD is the average log predictive density of the observed data points based on the leave-one-out posterior distributions. This measure has the advantage over other performance measures like classification accuracy of not only taking into account the location of the predictive distribution (the number of correct predictions) but also the width, i.e. how confident the model is in its predictions. For a description of ELPD in the context of LOOCV, see Vehtari *et al.* (2017) and Gneiting *et al.* (2007).

2.5 Data

The model was fit with several datasets consisting of viral sequences paired to host HLA class I data:

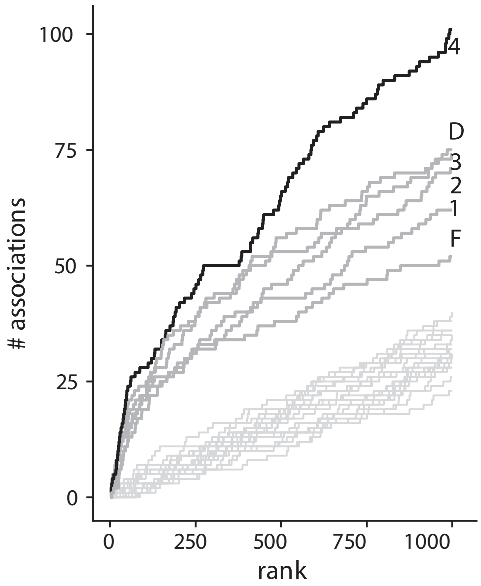


Fig. 1. HAM enrichment plot for HBV preC/core protein: number N_e of associations inside the boundary of known epitopes versus rank r. D: Phylogenetic Dependency Network; F: Fisher's exact test; 1: simple logistic regression model with broad Student-t priors; 2: logistic regression model with horseshoe prior; 3: logistic regression model with horseshoe prior and phylogeny; 4: full model with epitope prediction. Unannotated gray lines at the bottom of the graph are HAM enrichment curves for random permutations of the list of HLA allele—substitution pairs and act as baselines

- A large HIV dataset consisting of a subset of sequences from the HOMER (Brumme *et al.*, 2007, 2008a) cohort, the Western Australian HIV Cohort Study (WAHCS, Bhattacharya *et al.*, 2007; Moore, 2002) and participants of the US AIDS Clinical Trials Group (ACTG) protocol 5142 (John *et al.*, 2008) who also provided Human DNA under ACTG protocol 5128 (Haas *et al.*, 2003) (total $N = 1383$). These data were in part also used in the Phylogenetic Dependency Network study (Carlson *et al.*, 2008). The dataset contains sequences spanning the *gag*, *pol*, *env*, *nef*, *vif*, *vpr*, *vpu*, *tat* and *rev* genes.
- A set of 351 HIV sequences mostly spanning the *pol* gene from the Arevir database (Roomp *et al.*, 2006).
- A set of 544 Hepatitis-B-Virus sequences (Timm *et al.*, 2021). The dataset contains sequences of the preC/core, LHBs, Pol and HBx proteins.
- A set of 104 Hepatitis-D-Virus sequences containing the HDV-antigen (Karimzadeh *et al.*, 2018).
- A set of 41 HIV sequences spanning the *gag* and *pol* genes.

Lists of known epitopes were collected from the Immune Epitope Database (IEDB, Vita *et al.* (2019)). For HBV and HIV, we added data from the Hepitopes database (Lumley *et al.*, 2016) and the Los Alamos HIV Molecular Immunology Database (Yusim *et al.*, 2018), respectively. In total, we obtained 20 epitopes for HDV, 339 for HBV and 2684 epitopes for HIV. The counts refer to unique pairs of epitope and HLA alleles.

2.6 Data preparation

For all sequences, we applied the following preparation steps:

1. For each dataset, the sequences were split into subsequences, either by protein or gene.
2. If not already present in this format, sequences were translated into their amino acid representations.
3. Multiple sequence alignments were produced with MAFFT (Katoh *et al.*, 2013) (default parameters). In the few cases when the alignment introduced frameshifts, these were corrected manually.
4. RAxML-NG (Kozlov *et al.*, 2019) version 1.0.0 was used to generate a maximum likelihood phylogenetic tree for each gene/protein using the –model GTR+G+I option with all other parameters set to default values. If available, we used RNA or DNA sequences for this step, rather than protein sequences.

3 Results

In order to understand what the different building blocks of HAMdetector contribute, we applied four different Bayesian models of increasing complexity to each dataset, starting with the standard logistic regression model (Equation 1) and adding then the further components, i.e. the horseshoe prior (Equation 5), phylogeny (Equation 3) and epitope prediction, resulting in the full model (Supplementary Equation S1). For comparisons to existing methods, we also applied Fisher's exact test and the Phylogenetic Dependency Network Carlson *et al.* (2008) to the same data.

3.1 Run times and convergence

For a standard office computer, run times of HAMdetector on the smaller HDV dataset were of the order of minutes and on the order of hours for the Hepatitis B dataset. For the large HIV dataset, the models were run overnight. Run times scale approximately linearly with the product NK, where N is the number of sequences and K is the number of substitutions. All model fits showed no signs of inference issues. In total, samples were drawn from four Hamiltonian Markov chains with 1000 iterations each after 300 warm-up iterations. The effective sample size exceeded 200 for all model parameters, $\hat{R}$ convergence diagnostic values were below 1.1 in all cases.

3.2 Posterior predictive checks

The model yields well-calibrated posterior predictive probabilities of substitutions. This is exemplified in Supplementary Figure S1 for HBV core protein, but also holds true for the other datasets (Supplementary Fig. 'Calibration plots').

The predictions of the tested models are enriched in epitopes over baselines for almost all tested datasets (Fig. 1 for HBV preC/core protein and Supplementary Fig. 'HAM enrichment plots' for other datasets). Although the relative and absolute performance varies by protein (see Supplementary Fig. 'HAM enrichment summary'), HAMdetector consistently outperforms all other methods in all but two datasets, and performs on-par with the other methods in these two cases. For the best ranked HAMs, Fisher's exact test performs about as well as the HAMdetector backbone logistic regression model (model 1 in Fig. 1). Each of the following three model stages of HAMdetector increases HAM enrichment further. The horseshoe prior alone (model 2) is a drastic improvement over model 1, even though it does not include any specific external information. The logistic regression model with horseshoe prior works roughly as well as the Phylogenetic Dependency Network Carlson *et al.* (2008), which includes much more information. Model 3 with its additional inclusion of phylogeny has higher enrichment than model 2, and finally, the full model 4 with the inclusion of epitope prediction leads to a further improvement. Note that, model 4 only uses epitope *prediction* software and does not use any information of experimentally confirmed epitopes. The latter are here only used for model evaluation.

The Bayesian approach lends itself to incorporation of prior knowledge which usually helps in accurate modeling and prediction. In fact, a considerable effect is confirmed by the HAM enrichment plots with their ladder of improvements with increasing inclusion of information. It may be particularly surprising that the sparsifying horseshoe prior has such an impact although it does not use specific prior information. However, this is in principle the same mechanism as for the other information components: it is known that HAMs are sparse per HLA allele, and therefore supplying this information to the inference improves predictions. Figure 2 illustrates the effect of the sparsifying prior with an example, the substitution 11D in HIV integrase (Arevir dataset). There is no evidence for an association of HLA-A*01 with this substitution, whereas for HLA-B*44 the data is consistent with a strong association. The horseshoe prior has the effect of shrinking toward 0 specifically those regression coefficients with weak evidence of an association (A*01 in Fig. 2). This reduces the standard error for the remaining coefficients, leading in our example to narrowed histogram for the association with B*44 in the model with horseshoe prior.

3.3 Leave-one-out cross-validation

To quantify the ability of the four different model stages of HAMdetector to generalize to unseen cases, we computed the ELPD with Pareto-smoothed leave-one-out cross-validation. Table 1 shows results for the HBV preC/core protein in terms of ELPD changes with each new model stage. Each new model stage adds ELPD, i.e. is better at generalizing than the simpler model stages.

The model with horseshoe prior alone already has a much higher ELPD than the standard logistic regression model, even though it does not use any specific external data. This is because including the sparsity assumption allows the model to better separate signal from noise and the uncertainty of the close-to-zero coefficients does not propagate into uncertainty of predictions.

Including phylogeny further improves model performance a lot, as the assumption of independent and identically distributed data is replaced with specific information from the shared phylogenetic history.

While addition of sparsity and phylogeny has an effect on all substitutions and samples, epitope prediction only influences those substitutions that are restricted by a given HLA allele and only those samples that are annotated with the allele. Therefore, inclusion of epitope prediction does not improve ELPD as much as inclusion of phylogeny and the sparsity assumption. However, inclusion of epitope prediction is highly useful for determining which HLA alleles are associated with a substitution, as shown in the previous section.

Table 1. Prediction performance changes in terms of ELPD as HAMdetector components are added

	$\mathrm{ELPD}_{\mathrm{diff}}$	$\mathrm{se}_{\mathrm{diff}}$
Logistic regression (baseline)	0.0	0.0
+Horseshoe prior	949.8	65.2
+Phylogeny	4440.9	94.4
+Epitope prediction	63.1	18.9

Note: Values computed for HBV preC/core protein. Each value in the column $\mathrm{ELPD}_{\mathrm{diff}}$ is the ELPD difference to the model in the previous row, e.g. the ELPD difference between the model with epitope prediction and the previous one is 63.1. Models with larger ELPD have better predictive performance, e.g. the model with all components, including epitope prediction, has better predictive performance than the model lacking epitope prediction. All differences are several times the estimated standard error (column $\mathrm{se}_{\mathrm{diff}}$) away from zero, indicating that models that include more information have better predictive performance.

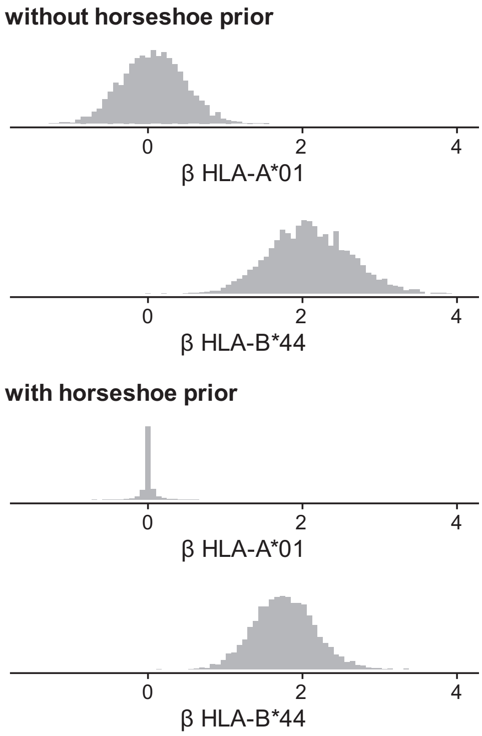


Fig. 2. Marginal posterior distributions of regression coefficients for the association of substitution 11D of the HIV integrase with HLA alleles A*01 and B*44. Top half: inferred with logistic regression model, bottom half: inferred with logistic regression with sparsifying horseshoe prior

3.4 HAMs in HDV as test case

The hepatitis D virus (HDV) dataset (Karimzadeh *et al.*, 2019) is an excellent test case: we have (i) a set of paired HDV sequences and patient HLA alleles, (ii) HAM predictions by Fisher's exact test as implemented in SeqFeatR (Budeus *et al.*, 2016) and (iii) an *in vitro* assay to quantify the effect of the predicted HAMs on IFN-γ release of CD8^{+} T cells (IFN-γ production assays, Karimzadeh *et al.*, 2019). This allows us to see whether HAMdetector decreases the false positive rate in comparison to the simpler Fisher's exact test, and we can make *bona fide* predictions on previously undetected HAMs. We have 15 HAMs predicted in HDV by Fisher's exact test at significance level 5×10^{-3} (Supplementary Table S1) as published (Karimzadeh *et al.*, 2019). The corresponding *P*-values have no clear relation to experimental confirmation, i.e. *P*-values for confirmed HAMs are not generally lower than those of non-confirmed ones.

For HAMdetector, we use in Supplementary Table S1 the posterior probability of a positive regression coefficient ($P(\beta_{jk} > 0)$ as measure for the confidence in having detected a HAM. HAMs with strong support have a posterior probability close to 1, associations with no support a probability close to 0.5 (corresponding to a regression coefficient centered around 0). The five predicted HAMs with top posterior probabilities (all ≥ 0.90) have all been experimentally confirmed. There is only one outlier with posterior probability 0.75 (P89T and B*37).

HAMdetector strongly supports 15 substitution—allele pairs that have previously not been identified (question marks in last column of Supplementary Table S1). All of them have association probabilities of 0.90 or higher, while their *P*-values from Fisher's exact test exceed the significance level of 5×10^{-3} used in Karimzadeh *et al.* (2019). Given the superior performance of HAMdetector on the experimentally tested HAMs, these 15 bona fide predictions suggest that most true HAMs may still to be discovered. A striking example is K43R—A*02 with a *P*-value of 0.22 in Fisher's exact test but a HAM-probability of 0.90 and location inside an A*02 restricted epitope.

3.5 Linkage disequilibrium

For three of the false positives proposed by Fisher's exact test (Supplementary Table S1), HAMdetector identifies associations with the same substitution but a different allele (P49L—B*13 instead of P49L—A*30; K43R—A*02 instead of K43R—B*13; and D33E—B*13 instead of D33E—A*03). One possible explanation for this observation is HLA linkage disequilibrium: If a certain HLA allele selects for a specific HAM and there is another HLA allele that co-occurs with that HLA allele, any method that relies on the statistical analysis of pairs of HLA allele and substitution alone will also detect these associations. Due to random sampling variation, the HLA allele that selects for a mutation might not necessarily have the strongest correlation. Inclusion of additional information like epitope prediction can help to identify associations that are otherwise confounded by noise.

Indeed, out of the 12 times P49L is observed in sequences annotated with A*30, B*13 is also present in 5 of those cases (Spearman's rank correlation coefficient $\rho = 0.5$). A similar observation can be made for K43R and D33E, although the correlation between the respective alleles is much weaker. A*30 and B*13 have been shown to be in strong linkage disequilibrium (Brumme *et al.*, 2007, Supplementary Table S2).

Figure 3 shows regression coefficients of the HLA alleles A*30 and B*13 for substitution P49L. With the simplest logistic regression model (model 1), both A*30 and B*13 have medium evidence of being associated with substitution P49L. However, with phylogeny and sparsity-promoting prior (model 3) both regression coefficients shrink close to 0—the associations are not convincingly supported by the data. Using epitope prediction as additional source of information (model 4) allows to disentangle the association of the correlated alleles with P49L and identify B*13 as likely associated with P49L. The association between P49L and A*30 (predicted by Fisher's exact test) remains shrunk toward 0.

3.6 HAMs outside epitopes

It is important to consider that biologically relevant HAMs do not necessarily have to lie within or close to the boundary of an epitope. In Supplementary Section S5, we outline that the model is still able to identify associations outside predicted epitopes and that epitope information augments evidence obtained from sequence data.

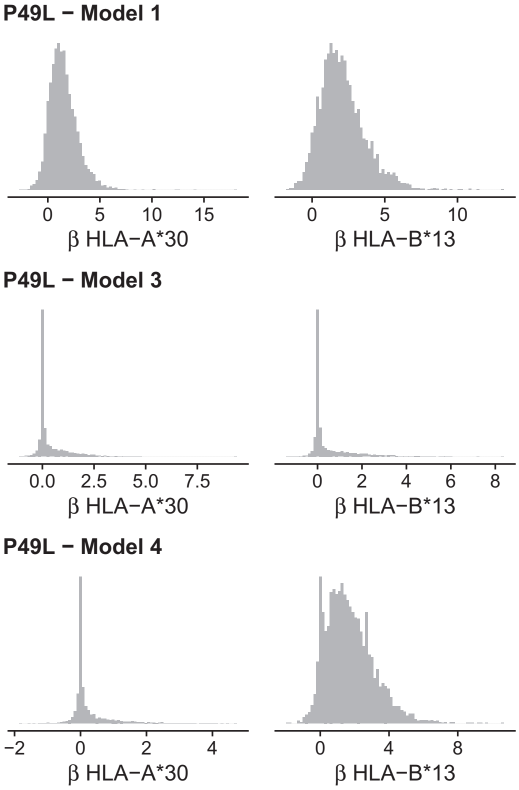


Fig. 3. Marginal posteriors for the regression coefficients of A*30 (left column) and B*13 (right column) for substitution P49L with different model stages (rows)

4 Discussion

HAMdetector follows a general paradigm of Bayesian modeling, namely to map all information that is available about a system of interest onto a probabilistic model, and then to apply Bayesian inference to learn about probable parameter values of that model, e.g. about β_{jk}, the association of HLA j with substitution k. The more relevant information we infuse into the model, the sharper the inference. HAMdetector outperforms other methods as it includes an unprecedented amount of relevant information.

We have demonstrated that the logistic regression backbone is a platform that can be extended by model components that contribute new information. We have selected such modules guided by widely accepted knowledge, such as phylogeny or epitope location. However, even knowledge that is rarely stated explicitly may be helpful in inference, as in the case of sparsity of HLA associations. Since the included knowledge is generic for interactions of variable viruses with CTL immunity, HAMdetector performance does not depend on the virus.

Yet, HAMdetector is far from perfect. For instance, the outlier in Supplementary Table S2 could point to missing information in HAMdetector. Another deficiency is that it currently works only with two-digit HLA alleles. We are currently exploring models for 4-digit HLA alleles that exploit partial pooling so that we can attenuate effects of the increased data fragmentation.

Generally, the platform character of HAMdetector model allows optimization of prediction performance by replacing components by more powerful ones, for example replacing a single epitope predictor by an ensemble predictor (Hu *et al.*, 2010). Another extension of our model would be to better account for phylogenetic uncertainty by using a Bayesian method to estimate a posterior distribution over possible tree topologies. The uncertainty over the tree topologies and the underlying parameters of the phylogenetic model would then propagate into uncertainty of the estimated probabilities $P(y_{ik} = 1|\Psi)$. However, the good performance of the current version of HAMdetector makes it already a valuable tool for the study of interactions between viruses and T-cell immunity.

Acknowledgements

The authors thank Drs. Mina John and Simon Mallal for providing data.

Funding

This work was supported by Deutsche Forschungsgemeinschaft (grant HO 1582/10-1). ZLB is supported by the Canadian Institutes for Health Research (through project grant PJT-148621) and by the Michael Smith Foundation for Health Research (through a Scholar Award).

Conflict of Interest: none declared.

References

Acevedo-Sáenz,L. *et al.* (2015) Selection pressure in CD8+ T-cell epitopes in the pol gene of HIV-1 infected individuals in Colombia. A bioinformatic approach. *Viruses*, 7, 1313–1331.

Alizon,S. *et al.* (2011) Epidemiological and clinical consequences of within-host evolution. *Trends Microbiol.*, **19**, 24–32.

Allen,T.M. *et al.* (2005) Selective escape from CD8+ T-cell responses represents a major driving force of human immunodeficiency virus type 1 (HIV-1) sequence diversity and reveals constraints on HIV-1 evolution. *J. Virol.*, **79**, 13239–13249.

Altman,J.D. *et al.* (1996) Phenotypic analysis of antigen-specific T lymphocytes. *Science*, **274**, 94–96.

Amrhein,V. *et al.* (2018) Remove, rather than redefine, statistical significance. *Nat. Hum. Behav.*, **2**, 4–4.

Baker,M. (2016) 1500 scientists lift the lid on reproducibility. *Nature*, **533**, 452–454.

Begley,C.G. *et al.* (2012) Drug development: raise standards for preclinical cancer research. *Nature*, **483**, 531–533.

Benjamini,Y. *et al.* (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodological)*, **57**, 289–300.

Betancourt, M. (2017). *A conceptual introduction to Hamiltonian Monte Carlo*. arXiv preprint arXiv:1701.02434.

Bezanson,J. *et al.* (2017) Julia: a fresh approach to numerical computing. *SIAM Rev.*, **59**, 65–98.

Bhattacharya,T. *et al.* (2007) Founder effects in the assessment of HIV polymorphisms and HLA allele associations. *Science*, **315**, 1583–1586.

Borrow,P. *et al.* (1997) Antiviral pressure exerted by HIV-l-specific cytotoxic T lymphocytes (CTLs) during primary infection demonstrated by rapid selection of CTL escape virus. *Nat. Med.*, **3**, 205–211.

Bronke,C. *et al.* (2013) HIV escape mutations occur preferentially at HLA-binding sites of CD8+ T-cell epitopes. *AIDS*, **27**, 899–905.

Brumme,Z.L. *et al.* (2007) Evidence of differential HLA class I-mediated viral evolution in functional and accessory/regulatory genes of HIV-1. *PLoS Pathogens*, **3**, e94.

Brumme,Z.L. *et al.* (2008a) Human leukocyte antigen-specific polymorphisms in HIV-1 Gag and their association with viral load in chronic untreated infection. *AIDS*, **22**, 1277–1286.

Brumme,Z.L. *et al.* (2008b) Marked epitope- and allele-specific differences in rates of mutation in human immunodeficiency type 1 (HIV-1) Gag, Pol, and Nef cytotoxic T-lymphocyte epitopes in acute/early HIV-1 infection. *J. Virol.*, **82**, 9216–9227.

Brunner,K.T. *et al.* (1968) Quantitative assay of the lytic action of immune lymphoid cells on 51-Cr-labelled allogeneic target cells *in vitro*; inhibition by isoantibody and by drugs. *Immunology*, **14**, 181–196.

Budeus,B. *et al.* (2016) SeqFeatR for the discovery of feature-sequence associations. *PLoS One*, **11**, e0146409.

Carlson,J.M. *et al.* (2008) Phylogenetic dependency networks: inferring patterns of CTL escape and codon covariation in HIV-1 Gag. *PLoS Comput. Biol.*, **4**, e1000225.

Carlson,J.M. *et al.* (2012) Widespread impact of HLA restriction on immune control and escape pathways of HIV-1. *J. Virol.*, **86**, 5230–5243.

Carvalho,C.M. *et al.* (2010) The horseshoe estimator for sparse signals. *Biometrika*, **97**, 465–480.

Czerkinsky,C.C. *et al.* (1983) A solid-phase enzyme-linked immunospot (ELISPOT) assay for enumeration of specific antibody-secreting cells. *J. Immunol. Methods*, **65**, 109–121.

Draenert,R. *et al.* (2004) Immune selection for altered antigen processing leads to cytotoxic T lymphocyte escape in chronic HIV-1 infection. *J. Exp. Med.*, **199**, 905–915.

Fisher,R.A. (1922) On the interpretation of X^2 from contingency tables, and the calculation of P. *J. R. Stat. Soc.*, **85**, 87.

Francke,U. *et al.* (1977) Assignment of the major histocompatibility complex to a region of the short arm of human chromosome 6. *Proc. Natl. Acad. Sci. USA*, **74**, 1147–1151.

Gabry,J. *et al.* (2019) Visualization in Bayesian workflow. *J. R. Stat. Soc. Ser. A (Stat. Soc.)*, **182**, 389–402.

Gelman,A. (2010) Bayesian statistics then and now. *Stat. Sci.*, **25**, 162–165.

Gelman,A. *et al.* (1992) Inference from iterative simulation using multiple sequences. *Stat. Sci.*, **7**, 457–472.

Gelman,A. *et al.* (2012) Why we (usually) don't have to worry about multiple comparisons. *J. Res. Educ. Effect.*, **5**, 189–211.

Gelman,A. *et al.* (2013) *Bayesian Data Analysis*. CRC Press, London.

Gelman,A. *et al.* (2014) Beyond power calculations. *Perspect. Psychol. Sci.*, **9**, 641–651.

Germain,R.N. (1994) MHC-dependent antigen processing and peptide presentation: providing ligands for T lymphocyte activation. *Cell*, **76**, 287–299.

Gneiting,T. *et al.* (2007) Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.*, **102**, 359–378.

Goldberg,A.L. *et al.* (2002) The importance of the proteasome and subsequent proteolytic steps in the generation of antigenic peptides. *Mol. Immunol.*, **39**, 147–164.

Haas,D.W. *et al.*; Adult AIDS Clinical Trials Group. (2003) A multi-investigator/institutional DNA bank for AIDS-related human genetic studies: AACTG Protocol A5128. *HIV Clin. Trials*, **4**, 287–300.

Harty,J.T. *et al.* (2000) CD8+ T cell effector mechanisms in resistance to infection. *Annu. Rev. Immunol.*, **18**, 275–308.

Hu,X. *et al.* (2010) MetaMHC: a meta approach to predict peptides binding to MHC molecules. *Nucleic Acids Res.*, **38**, W474–W479.

Ioannidis,J.P.A. (2005) Why most published research findings are false. *PLoS Med.*, **2**, e124.

Ives,A.R. *et al.* (2010) Phylogenetic logistic regression for binary dependent variables. *Syst. Biol.*, **59**, 9–26.

Ives,A.R. *et al.* (2014) Phylogenetic regression for binary dependent variables. In: Garamszegi, L.Z. (ed.) *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology*. Springer, Berlin, Heidelberg, pp. 231–261.

John,M. *et al.* (2008) Genome-wide HLA-associated selection in HIV-1 and protein-specific correlations with viral load: an ACTG5142. In: *15th Conference on Retroviruses and Opportunistic Infections (CROI) (Abstract 312), Boston, MA*.

Karimzadeh,H. *et al.* (2018) Amino acid substitutions within HLA-B27-restricted T cell epitopes prevent recognition by hepatitis delta virus-specific CD8+ T cells. *J. Virol.*, **92**, e01891-17.

Karimzadeh,H. *et al.* (2019) Mutations in hepatitis D virus allow it to escape detection by CD8+ T cells and evolve at the population level. *Gastroenterology*, **156**, 1820–1833.

Katoh,K. *et al.* (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.

Kawashima,Y. *et al.* (2009) Adaptation of HIV-1 to human leukocyte antigen class I. *Nature*, **458**, 641–645.

Kløverpris,H.N. *et al.* (2015) Role of HLA adaptation in HIV Evolution. *Front. Immunol.*, **6**, 665.

Kozlov,A.M. *et al.* (2019) RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, **35**, 4453–4455.

Lamoreaux,L. *et al.* (2006) Intracellular cytokine optimization and standard operating procedure. *Nat. Protoc.*, **1**, 1507–1516.

Le Gall,S. *et al.* (2007) Portable flanking sequences modulate CTL epitope processing. *J. Clin. Investig.*, **117**, 3563–3575.

Leslie,A.J. *et al.* (2004) HIV evolution: CTL escape mutation and reversion after transmission. *Nat. Med.*, **10**, 282–289.

Lumley,S. *et al.* (2016) Hepitopes: a live interactive database of HLA class I epitopes in hepatitis B virus. *Wellcome Open Res.*, **1**, 9.

Lumley,S.F. *et al.* (2018) Hepatitis B virus adaptation to the CD8+ T cell response: consequences for host and pathogen. *Front. Immunol.*, **9**, 1561.

Matthews,P.C. *et al.* (2008) Central role of reverting mutations in HLA associations with human immunodeficiency virus set point. *J. Virol.*, **82**, 8548–8559.

Mei,S. *et al.* (2020) A comprehensive review and performance evaluation of bioinformatics tools for HLA class I peptide-binding prediction. *Brief. Bioinf.*, **21**, 1119–1135.

Milicic,A. *et al.* (2005) CD8+ T cell epitope-flanking mutations disrupt proteasomal processing of HIV-1 Nef. *J. Immunol.*, **175**, 4618–4626.

Moore,C.B. *et al.* (2002) Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. *Science*, **296**, 1439–1443.

Murata,S. *et al.* (2007) Regulation of CD8+ T cell development by thymus-specific proteasomes. *Science*, **316**, 1349–1353.

O'Donnell,T.J. *et al.* (2020) MHCflurry 2.0: improved pan-allele prediction of MHC class I-presented peptides by incorporating antigen processing. *Cell Syst.*, **11**, 42–48.e7.

Osborne,J.W. *et al.* (2002) Four assumptions of multiple regression that researchers should always test. Pract. Assess. Res. Eval., 8, Article 2.

Piironen,J. *et al.* (2017a) On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (pp. 905-913). PMLR.

Piironen,J. *et al.* (2017b) Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electron. J. Stat.*, **11**, 5018–5051.

Robinson,J. *et al.* (2015) The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res.*, **43**, D423–D431.

Roomp,K. *et al.* (2006) Arevir: a secure platform for designing personalized antiretroviral therapies against HIV. In: Leser, U., Naumann, F. and Eckman, B. (eds.), *Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg, pp. 185–194.

Rousseau,C.M. *et al.* (2008) HLA class I-driven evolution of human immunodeficiency virus type 1 subtype C proteome: immune escape and viral load. *J. Virol.*, **82**, 6434–6446.

Ruhl,M. *et al.* (2012) Escape from a dominant HLA-B15-restricted CD8 T cell response against hepatitis C virus requires compensatory mutations outside the epitope. *J. Virol.*, **86**, 991–1000.

Scariano,S.M. *et al.* (1987) The effects of violations of independence assumptions in the one-way ANOVA. *Am. Stat.*, **41**, 123–129.

Stan Development Team. (2021) Stan Modeling Language Users Guide and Reference Manual, 2.23. https://mc-stan.org/users/citations/.

Timm,J. *et al.* (2021) GenBank Accession Numbers: Genotype A: MZ043025 – MZ043097, Genotype B: MW845286 – MW845312, Genotype C: MW887641 – MW887652, Genotype D: MZ097624 – MZ097884, Genotype E: MW926548 – MW926566.

Vehtari,A. *et al.* (2017) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.*, **27**, 1413–1432.

Vita,R. *et al.* (2019) The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.*, **47**, D339–D343.

Yewdell,J.W. *et al.* (2002) Viral interference with antigen presentation. *Nat. Immunol.*, **3**, 1019–1025.

Yokomaku,Y. *et al.* (2004) Impaired processing and presentation of cytotoxic-T-lymphocyte (CTL) epitopes are major escape mechanisms from CTL immune pressure in human immunodeficiency virus type 1 infection. *J. Virol.*, **78**, 1324–1332.

Yusim,K. *et al.* (eds.) (2018) *HIV Molecular Immunology*. Los Alamos National Laboratory, Theoretical Biology and Biophysics, Los Alamos, New Mexico.

HAMdetector: A Bayesian regression model that integrates information to detect HLA-associated mutations

Supplementary Information

Daniel Habermann[1,*], Hadi Kharimzadeh[2], Andreas Walker[3], Yang Li[4], Rongge Yang[4], Zabrina L. Brumme[5,6], Jörg Timm[3], Michael Roggendorf[7], and Daniel Hoffmann[1,8,9]

[1] *Bioinformatics and Computational Biophysics, Faculty of Biology, University of Duisburg-Essen, Essen, 45117, Germany*
[2] *Division of Clinical Pharmacology, University Hospital, LMU Munich, Munich, Germany*
[3] *Institute of Virology, Medical Faculty, University Hospital Düsseldorf, Heinrich-Heine-Universität, Düsseldorf, 40225, Germany*
[4] *AIDS and HIV Research Group, State Key Laboratory of Virology, Wuhan Institute of Virology, Chinese Academy of Science, Wuhan, P. R. China*
[5] *Faculty of Health Sciences, Simon Fraser University, Burnaby, Canada*
[6] *British Columbia Centre for Excellence in HIV/AIDS, Vancouver, Canada*
[7] *Insitute of Virology, School of Medicine, Technical University of Munich/Helmholtz Zentrum München, Munich, Germany*
[8] *Center of Medical Biotechnology, University of Duisburg-Essen, Essen, Germany*
[9] *Center for Computational Sciences and Simulation, University of Duisburg-Essen, Essen, Germany*

[*] To whom correspondence should be addressed.

Contents

1 Full model specification

The full model specification is given by:

$$
\begin{aligned}
y_{ik} &\sim \text{Bernoulli}(\theta_{ik}) \\
\theta_{ik} &= \text{logistic}\Big(\beta_{0_k} + \gamma_k \text{logit}\left(P(y_{ik} = 1|\Psi)\right) \\
&\quad + \sum_{k=1}^{D} X_{ij}\beta_{jk}\Big) \\
\beta_{0_k} &\sim \text{Normal}(0, 100^2) && (*) \\
\gamma_k &\sim \text{Normal}(\mu_{\text{phy}}, \sigma^2_{\text{phy}}) \\
\mu_{\text{phy}} &\sim \text{Normal}(1, 1) && (*) \\
\sigma_{\text{phy}} &\sim \text{Normal}^+(0, 0.5) && (*) \\
\beta_{\text{epi}} &\sim \text{Normal}^+(1, 2) && (*) \\
\beta_{jk} &\sim \text{Normal}(0, \tau_k^2 \tilde{\lambda}_{jk}^2) \\
\tilde{\lambda}_{jk}^2 &= \frac{c_k^2 \lambda_{jk}^2}{c_k^2 + \tau_k^2 \lambda_{jk}^2} \\
c_k^2 &\sim \text{Inv-Gamma}(3.5, 3.5) && (*) \\
\lambda_{jk} &\sim \text{Cauchy}^+(0, \sigma_j \text{exp}(Z_{jk}\beta_{\text{epi}})) \\
\tau_k &\sim \text{Cauchy}^+(0, \tau_{0k}) && (*) \\
\tau_{0k} &= \frac{10}{D - 10}\frac{2}{\sqrt{N}}
\end{aligned} \tag{1}
$$

- y_{ik} is the binary encoded observation of substitution k in viral sequence i (each observed amino acid state k contributes a separate column to y_{ik}).
- θ_{ik} is the estimated probability that we observe substitution k in sequence i.

- β_{0k} is an intercept for substitution k, corresponding to the overall log-odds for substitution k in absence of any HLA-specific effects.

- X_{ij} is 1 if sequence i comes from host individual with HLA allele j and 0 otherwise.

- β_{jk} is the HLA regression coefficient of HLA allele j for substitution k.

- D is the number of HLA alleles in the dataset.

- the logistic inverse link function transforms the linear model in parentheses to the probability scale of θ_{ik}.

- γ_k are (hierarchical) phylogeny coefficients accounting for uncertainty of the phylogenetic model.

- β_{epi} governs the increase in scale of the corresponding local shrinkage parameters, which means that HLA regression coefficients for positions inside predicted epitopes are more likely to be non-zero.

- β_{jk} is the HLA regression coefficient of HLA allele j for substitution k.

- c governs the magnitude of shrinkage for the non-zero coefficients

- τ and λ_{jk} are the so-called global and local shrinkage parameters, respectively.

- Cauchy$^+$ is the positively constrained Cauchy distribution.

- τ_0 is the overall degree of sparsity.

- N is the total number of annotated sequences.

The full model specification includes some aspects that were not covered in the main manuscript. In particular, the overall phylogeny-weight γ in Eq 1 of the manuscript is replaced by hierarchically modelled γ_k, which allows partial pooling across substitutions (even with a global parameter γ the model works reasonably well). The final additional parameters $\tilde{\lambda}_{jk}^2$ and τ_{0k} are explained in detail in [1]. Briefly, $\tilde{\lambda}_{jk}^2$ allows some regularization for the non-zero coefficients and the parameterization of τ_{0k} allows to place a prior on the expected number of non-zero coefficients. This is particularly useful for logistic regression models, as some shrinkage helps to deal with issues of separability and collinearity that commonly occur with logistic regression models.

2 Prior justification

Prior distributions are labeled with an asterisk in Eq 1. They are weakly informative, which means that they effectively limit posteriors to realistic magnitudes of parameters. One exception to this are the intercepts β_{0_k}, which are essentially flat because they are well identified by the data alone.

The hierarchical mean and standard deviation of the phylogeny coefficients γ_k place most probability mass on γ_k values around 1. In absence of any HLA effects, a $\gamma_k = 1$ would mean that the estimate for the probability of observing substitution k is identical to the probability based on the phylogenetic model. This treats phylogeny as a baseline, and any observations not attributed to phylogeny must be explained by HLA alleles or noise.

The prior on c_k^2 implies a Student-t prior with 7 degrees of freedom and a scale of 1 on the non-zero HLA regression coefficients β_{jk}. A Student-t prior with these parameters is a reasonable default choice for logistic regression models [2].

The value of τ_{0k} implies 10 effective non-zero HLA regression coefficients per substitution. The rationale behind this parameterization is again outlined in [1]. The value of 10 corresponds to a generously estimated magnitude based on available HIV epitope maps [3]. The model is also parameterized in a way that assumes an equal degree of sparsity across all alignment positions a priori. We also tried to model τ_k hierarchically, but observed sampling issues due to the resulting unfavorable geometry of the posterior.

3 HBV PreC/core calibration plot

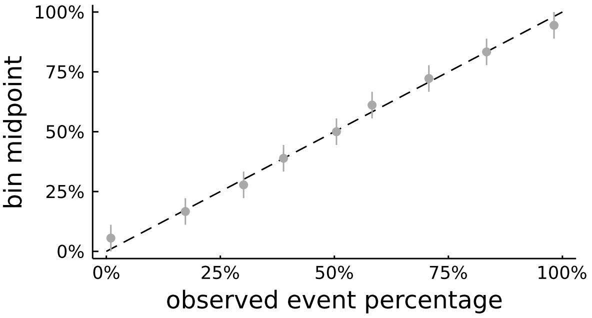


Figure 1: Calibration plot for the HBV PreC/core protein.

4 HAMs in HDV as a test case

Table 1: List of HAMs predicted by Fisher's exact test (FET). In the last column "+" and "-" mark experimentally confirmed or rejected HAMs, respectively; "?" below the horizontal line indicate untested bona fide predictions. "post.prob." are posterior probabilities for positive associations computed with HAMdetector.

substitution	allele	p-value (FET)	post. prob.	confirmed
S170N	B*15	$3 \cdot 10^{-8}$	0.99	+
D101E	B*37	0.0002	0.96	+
R105K	B*27	0.0011	0.93	+
R139K	B*41	0.0034	0.92	+
E47D	B*18	0.0027	0.90	+
D33E	B*13	0.0001	0.86	-
T134A	A*68	0.0045	0.82	-
K43R	B*13	0.0021	0.77	-
P89T	B*37	0.0011	0.75	+
D47E	A*30	0.0010	0.76	-
K113R	B*13	0.0043	0.76	-
A107T	B*14	0.0028	0.70	-
P49L	A*30	0.0031	0.63	-
Q100L	B*13	0.0018	0.60	-
D96E	B*13	0.0035	0.51	-
E46D	A*02	0.0054	0.97	?
V81I	A*68	0.0073	0.97	?
K113N	B*08	0.0063	0.96	?
A71T	B*41	0.0065	0.96	?
L188I	A*68	0.0632	0.94	?
T95S	A*01	0.0285	0.93	?
D33E	A*03	0.0226	0.93	?
P49L	B*13	0.0035	0.92	?
A74S	A*68	0.0123	0.91	?
E29D	B*44	0.0559	0.91	?
D46E	B*57	0.0190	0.91	?
R88K	A*68	0.0123	0.91	?
T149P	B*52	0.0281	0.91	?
K43R	A*02	0.2158	0.90	?
N22S	B*08	0.0405	0.90	?

5 HAMs outside epitopes

The epitope and processing predictions that HAMdetector uses are imperfect, as the underlying tools extrapolate binding affinities for new epitopes based on necessarily incomplete experimental data. Bayesian statistics provides a coherent framework to make use of imperfect data. In HAMdetector, this is achieved by an additional parameter β_{epi} that governs how strongly the model takes an apparent association between a substitution and the corresponding HLA allele into account. By default, the regression coefficients that quantify the strength of association between allele and substitution are shrunk towards 0, and only in the presence of considerable evidence in favor of an association (e.g. because the substitution often co-occurs with a certain HLA allele), this shrinkage is overcome by the observed data.

If the epitope prediction happens to be reliable, i.e. when the presence of a predicted epitope correlates strongly with the probability observing the substitution in a host with the respective HLA allele, the parameter β_{epi} is estimated to be large and less evidence by the sequence data is enough to escape the shrinkage and estimate a non-zero association between allele and substitution, compared to associations that do not lie inside a predicted epitope. Likewise, if the epitope prediction turns out to be non-reliable, β_{epi} is estimated to be close to 0 and the presence of a predicted epitope does not strongly affect the conclusions drawn from the sequence data.

However, it is important to consider that biologically relevant HAMs do not necessarily have to lie within or close to the boundary of an epitope. For instance, compensatory mutations can occur far away from the epitope they are associated with, as they might be the result of improved physical interactions with another amino acid in the folded, three-dimensional protein [4]. Such compensatory mutations [5, 4, 6, 7] can confer a strong selection advantage, e.g. by partially restoring replicative capacity that would otherwise be impaired by the exclusive presence of a certain HLA escape mutation.

We therefore also expect HAMs outside epitopes and one possible concern is that the model focuses too strongly on associations with substitutions that lie within the boundary of predicted epitopes.

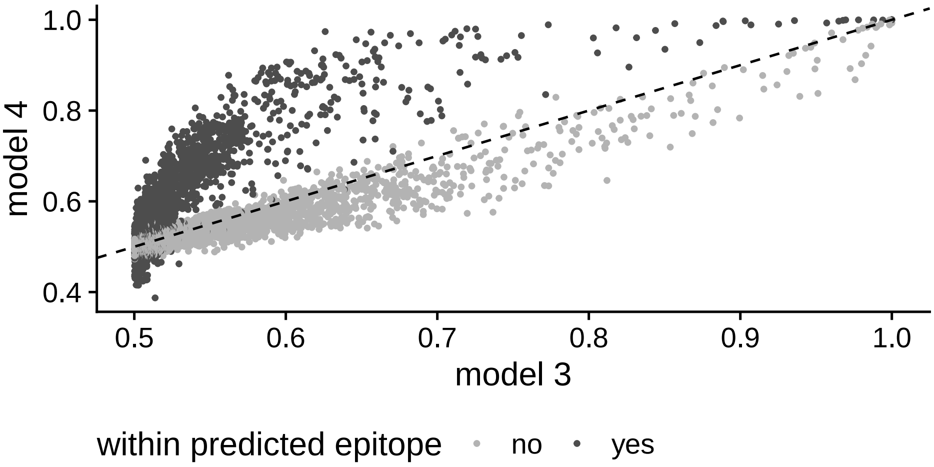


Figure 2: Integral of the marginal posterior $P(\beta_{jk} > 0)$ for the HAMdetector model with epitope prediction (model 4) and without epitope prediction (model 3) for all substitutions in the preC/core protein (HBV dataset).

Figure 2 shows posterior probabilities $P(\beta_{jk} > 0)$ for substitution–HLA allele pairs as calculated by HAMdetector with (model 4) and without (model 3) epitope prediction. Each substitution–HLA allele pair is represented by a dot and colored according to whether or not that position lies within a predicted epitope. For substitutions that do not lie within a predicted epitope, both models provide similar estimates (points along the diagonal). However, some substitutions–HLA pairs that have only weak evidence of association in model 3 have strong support in model 4, which is explained by the additional evidence provided by epitope prediction. The figure shows that the model is still able to identify associations outside predicted epitopes and that epitope information augments evidence obtained from sequence data.

References

[1] Juho Piironen and Aki Vehtari. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2):5018–5051, 2017.

[2] Stan prior choice recommendations. `https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations/ccced280cdb53a195fcd2b26168587fe4a2a4d27`. Accessed: 2021-19-07.

[3] Karina Yusim, Elizabeth-Sharon David-Fung, Bette T. M. Korber, Christian Brander, Dan Barouch, Rob de Boer, Barton F. Haynes, Richard Koup, John P. Moore, Bruce D. Walker, and David I. Watkins, editors. *HIV Molecular Immunology.* Los Alamos National Laboratory, Theoretical Biology and Biophysics, Los Alamos, New Mexico, 2018.

[4] M. Ruhl, P. Chhatwal, H. Strathmann, T. Kuntzen, D. Bankwitz, K. Skibbe, A. Walker, F. M. Heinemann, P. A. Horn, T. M. Allen, D. Hoffmann, T. Pietschmann, and J. Timm. Escape from a dominant HLA-B*15-restricted CD8 T Cell response against hepatitis C virus requires compensatory mutations outside the epitope. *Journal of Virology*, 86(2):991–1000, nov 2011.

[5] Anthony D. Kelleher, Chad Long, Edward C. Holmes, Rachel L. Allen, Jamie Wilson, Christopher Conlon, Cassy Workman, Sunil Shaunak, Kara Olson, Philip Goulder, Christian Brander, Graham Ogg, John S. Sullivan, Wayne Dyer, Ian Jones, Andrew J. McMichael, Sarah Rowland-Jones, and Rodney E. Phillips. Clustered mutations in HIV-1 gag are consistently required for escape from HLA-B27–restricted cytotoxic T lymphocyte responses. *Journal of Experimental Medicine*, 193(3):375–386, feb 2001.

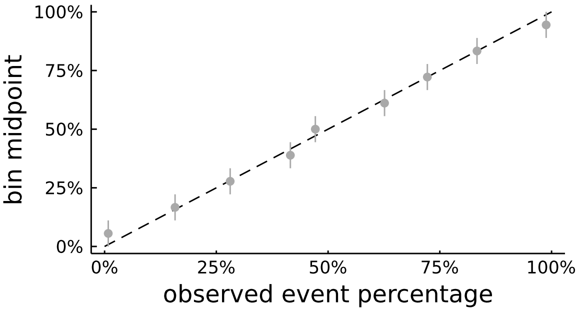
[6] Christoph Neumann-Haefelin, Cesar Oniangue-Ndza, Thomas Kuntzen, Julia Schmidt, Katja Nitschke, John Sidney, Célia Caillet-Saguy, Marco Binder, Nadine Kersting, Michael W. Kemper, Karen A. Power, Susan Ingber, Laura L. Reyor, Kelsey Hills-Evans, Arthur Y. Kim, Georg M. Lauer, Volker Lohmann, Alessandro Sette, Matthew R. Henn, Stéphane Bressanelli, Robert Thimme, and Todd M. Allen. Human leukocyte antigen B27 selects for rare escape mutations that significantly impair hepatitis C virus replication and require compensatory mutations. *Hepatology*, 54(4):1157–1166, sep 2011.

[7] Arne Schneidewind, Mark A. Brockman, John Sidney, Yaoyu E. Wang, Huabiao Chen, Todd J. Suscovich, Bin Li, Rahma I. Adam, Rachel L. Allgaier, Bianca R. Mothé, Thomas Kuntzen, Cesar Oniangue-Ndza, Alicja Trocha, Xu G. Yu, Christian Brander, Alessandro Sette, Bruce D. Walker, and Todd M. Allen. Structural and functional constraints limit options for cytotoxic t-lymphocyte escape in the immunodominant HLA-b27-restricted epitope in human immunodeficiency virus type 1 capsid. *Journal of Virology*, 82(11):5594–5605, jun 2008.

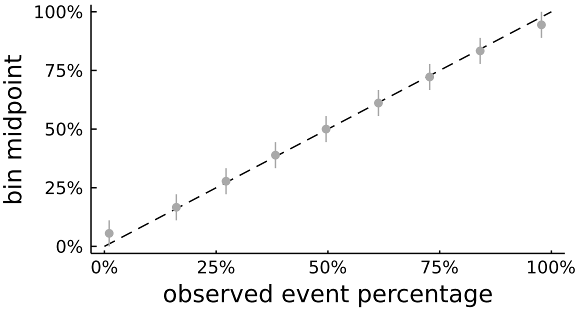
6 Calibration plots

HIV HOMER+

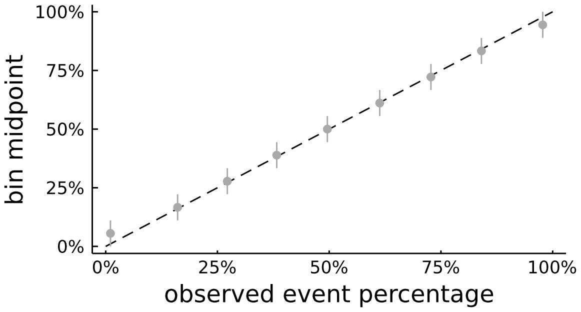
gag



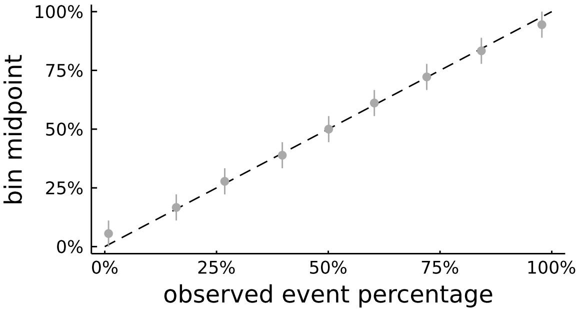
gp41



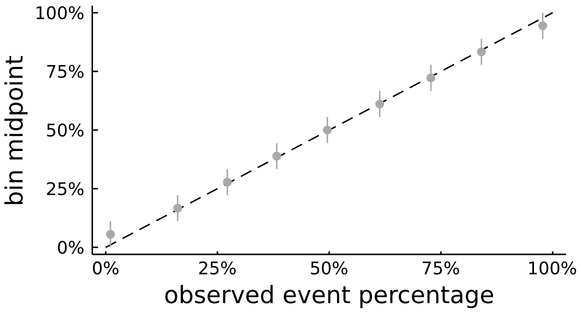
gp120



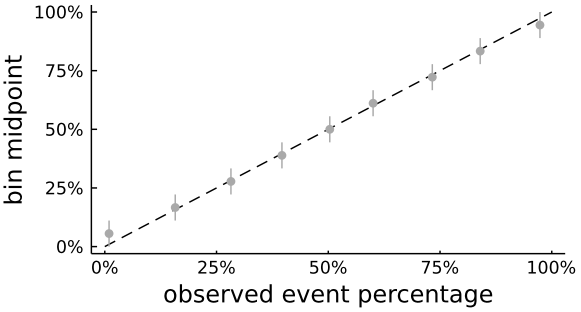
nef



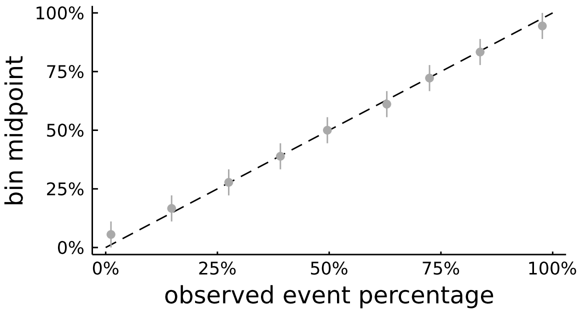
pol



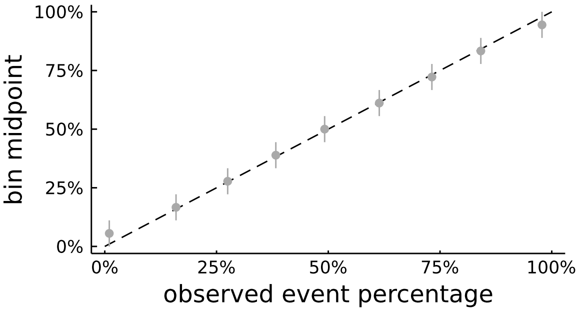
rev



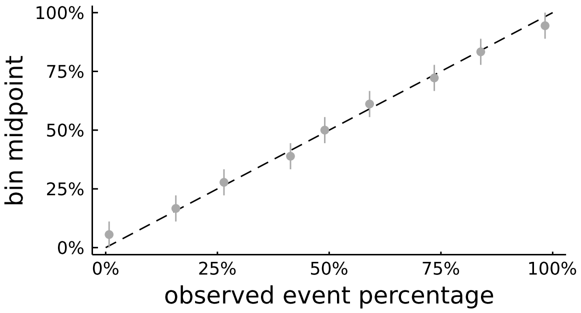
tat



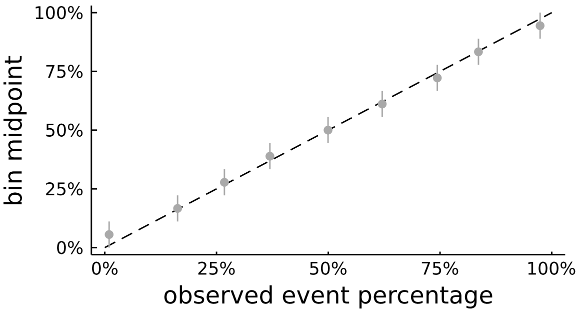
vif



vpr

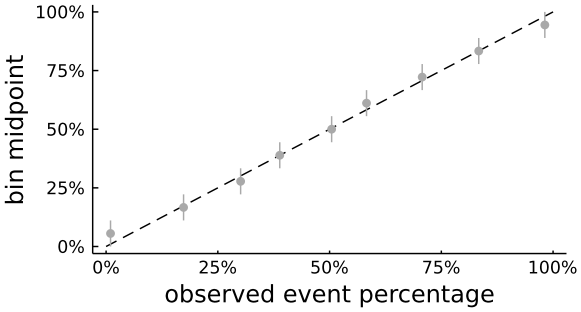

bin midpoint
100%
75%
50%
25%
0%
0%
25%
50%
75%
100%
observed event percentage

vpu

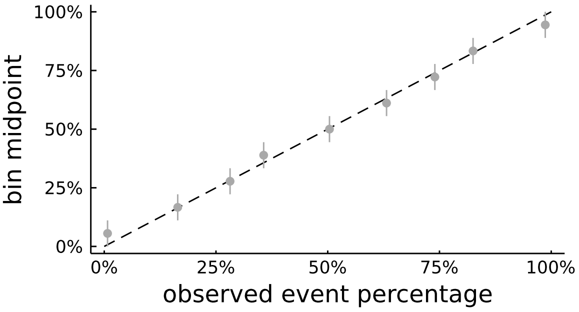

bin midpoint
100%
75%
50%
25%
0%
0%
25%
50%
75%
100%
observed event percentage

HBV

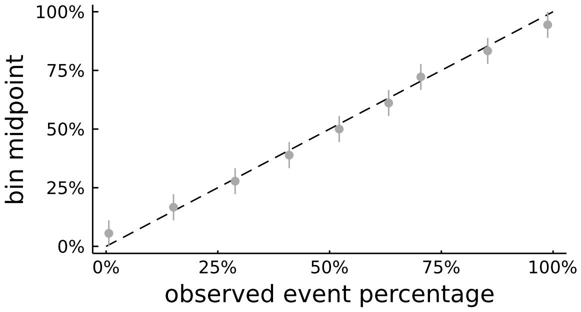
preC/core



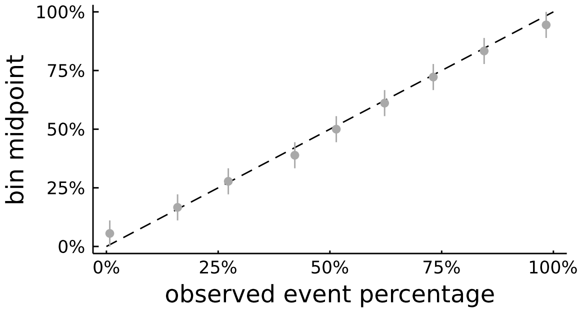
HBx



LHBs

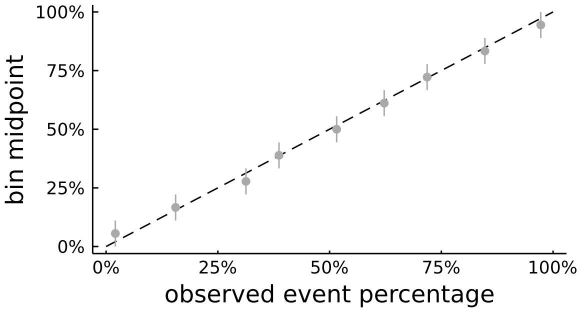


Pol



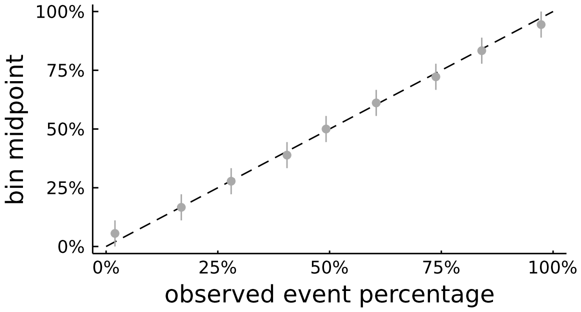
HDV

delta

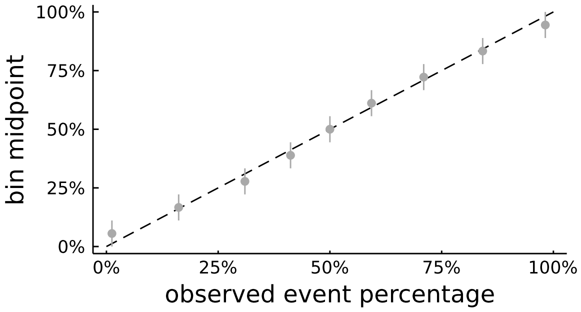


HIV Arevir

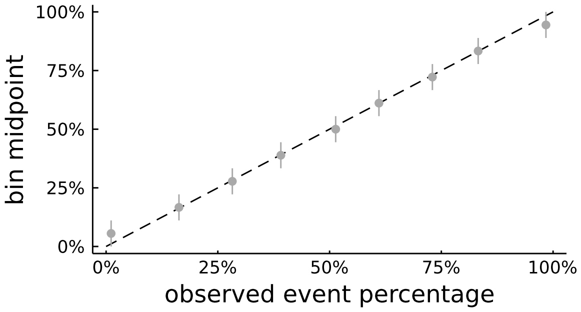
gp120



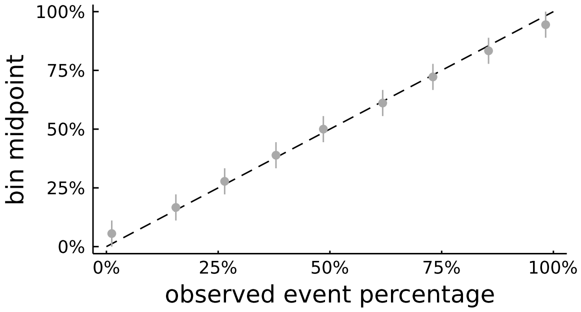
integrase



protease

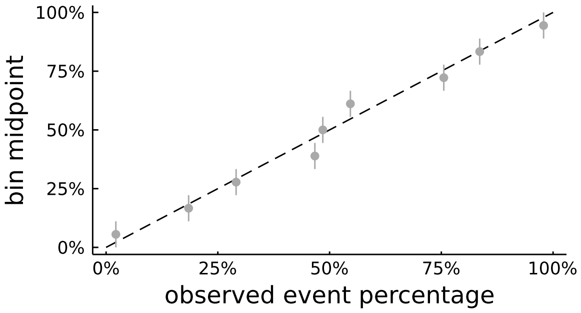


reverse transcriptase

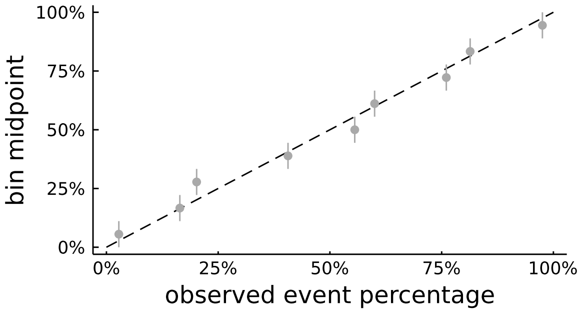


HIV

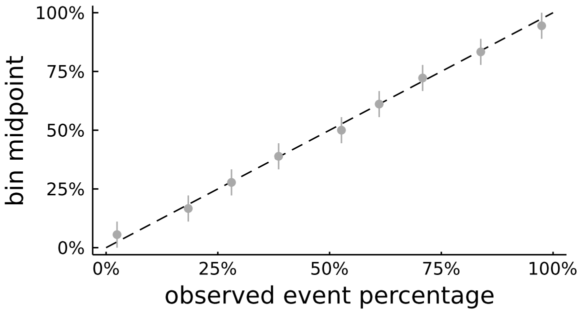
gag



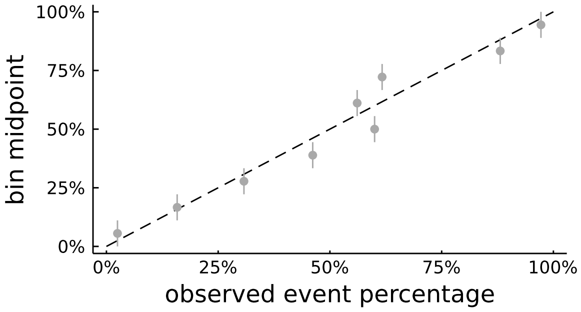
integrase



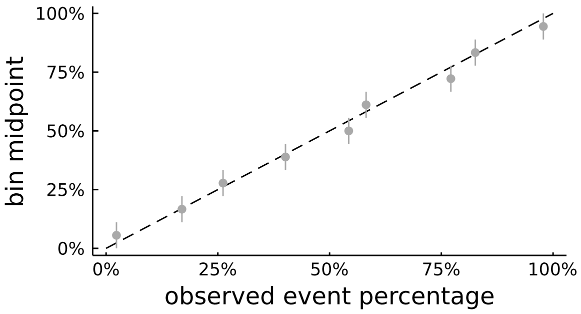
pol



protease



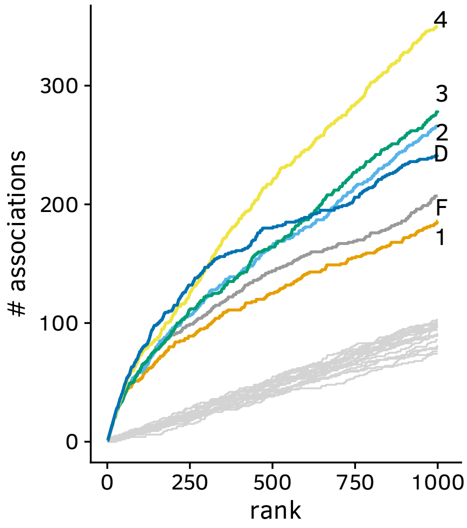
reverse transcriptase



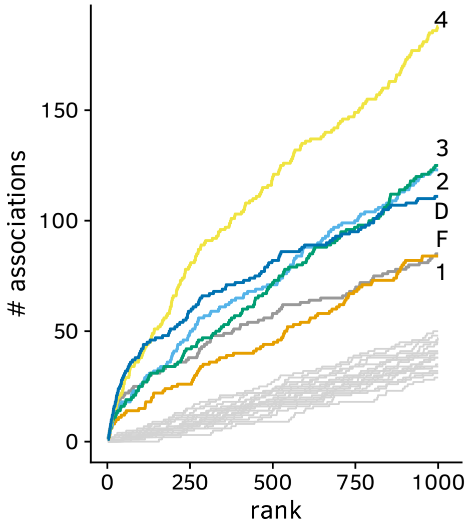
7 HAM enrichment plots

HIV HOMER+

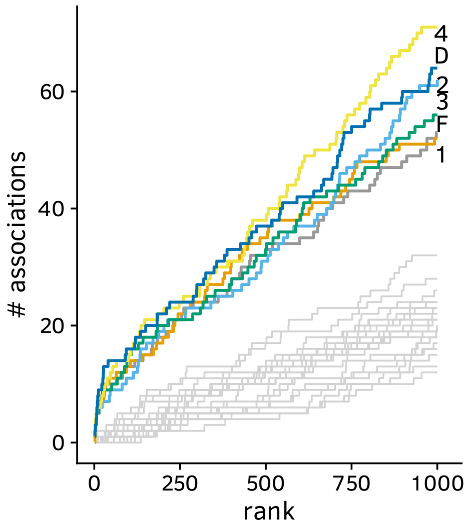
gag



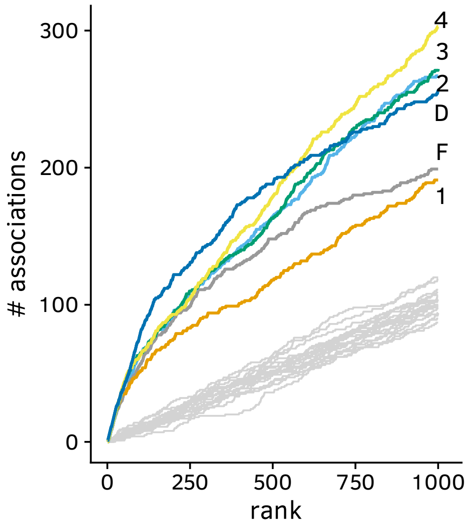
gp41



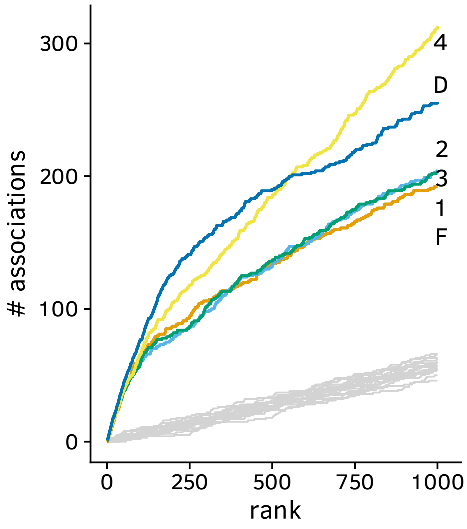
gp120



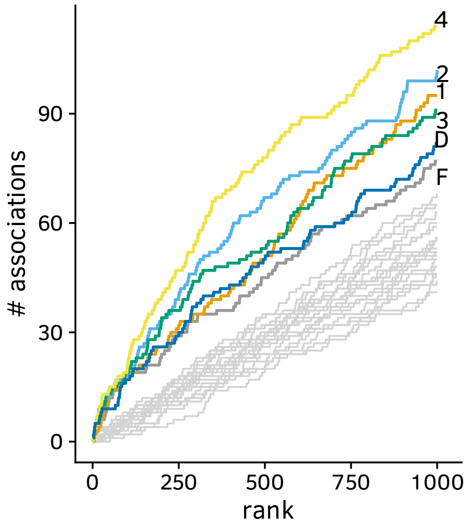
nef

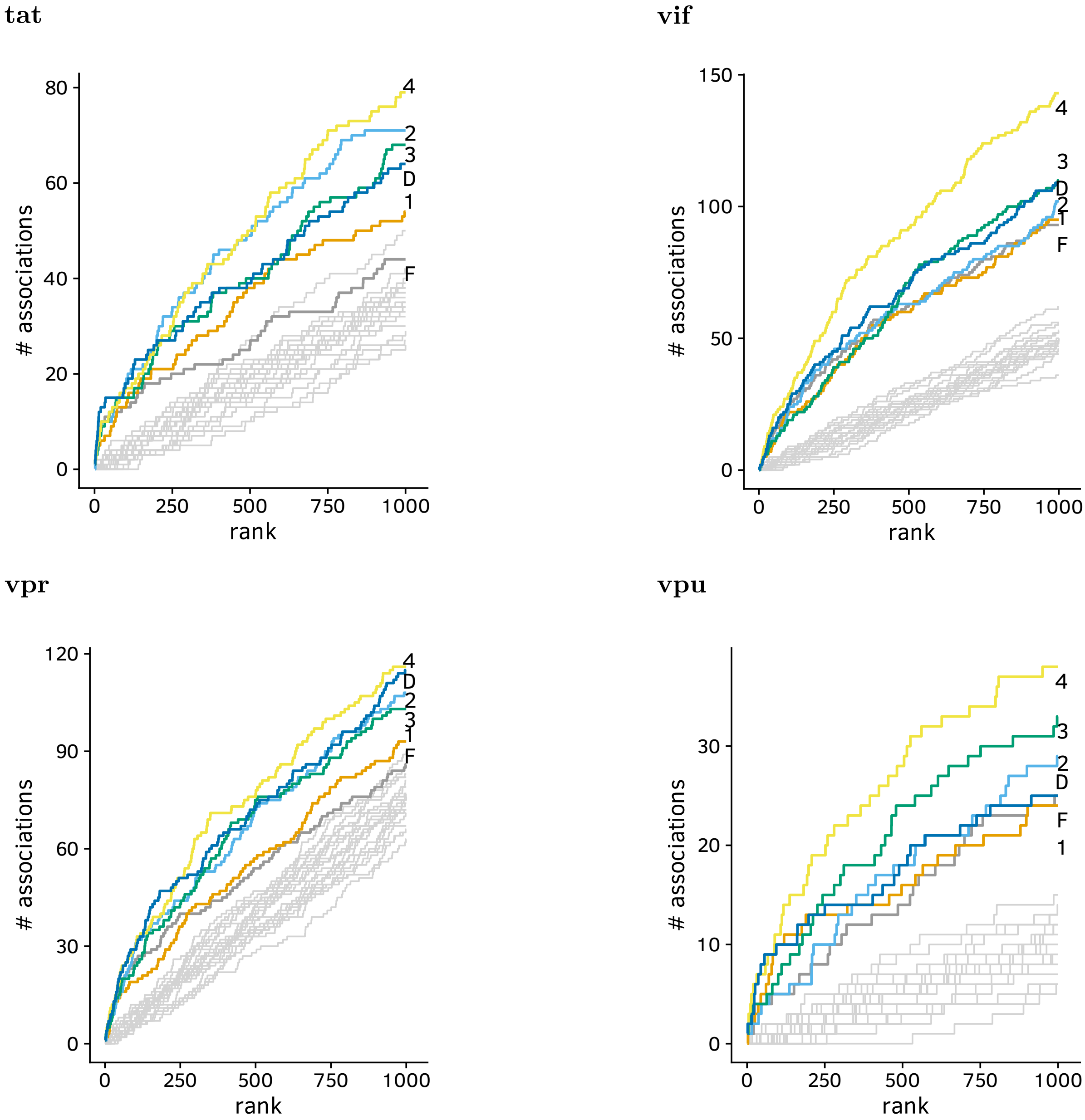


pol



rev





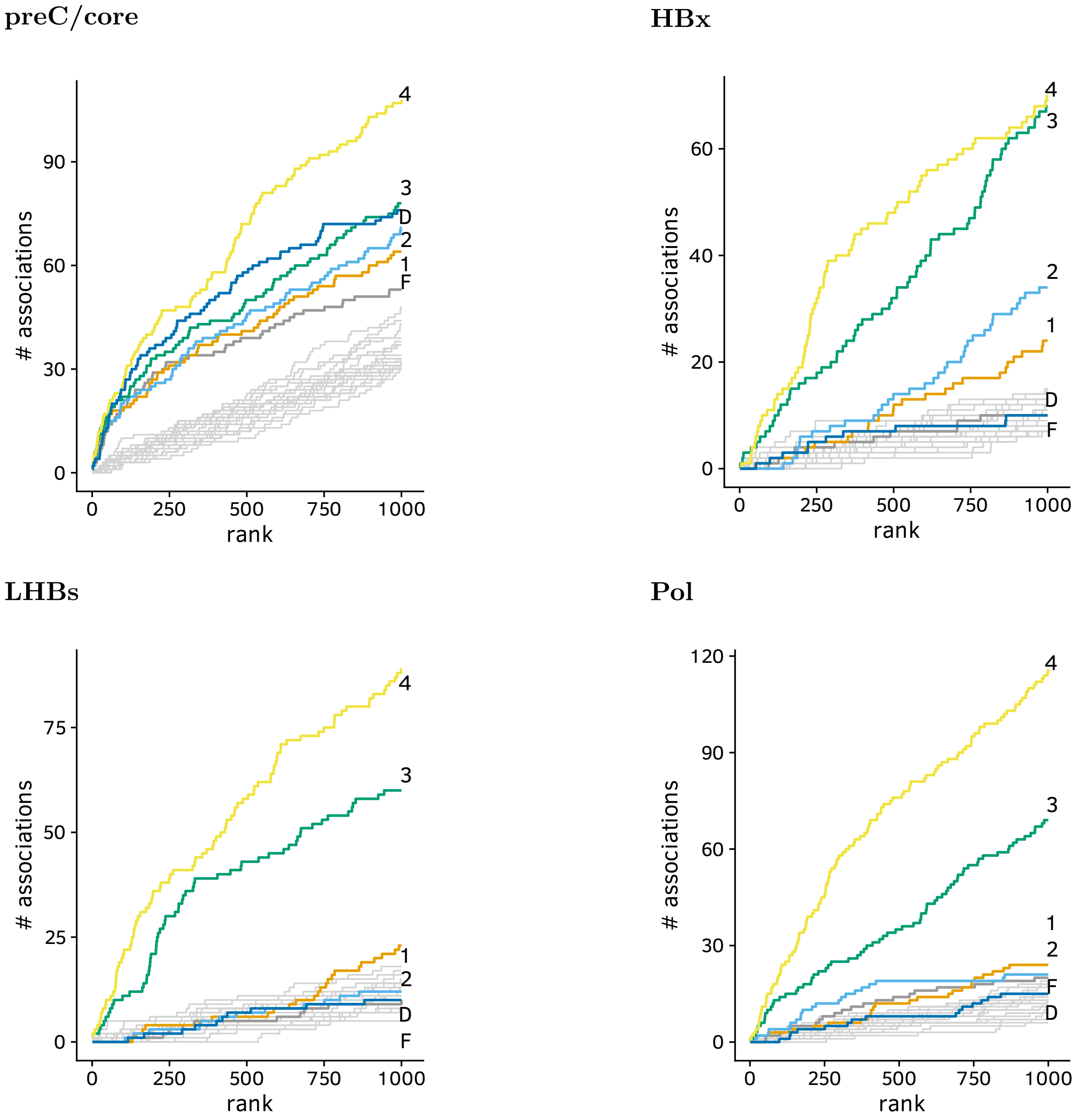
tat

vif

vpr

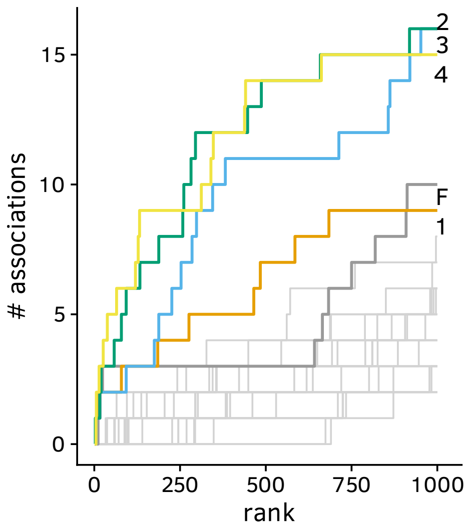
vpu

HBV


preC/core
# associations
0
30
60
90
rank
0
250
500
750
1000
4
3
D
2
1
F
HBx
# associations
0
20
40
60
rank
0
250
500
750
1000
4
3
2
1
D
F
LHBs
# associations
0
25
50
75
rank
0
250
500
750
1000
4
3
1
2
D
F
Pol
# associations
0
30
60
90
120
rank
0
250
500
750
1000
4
3
1
2
F
D

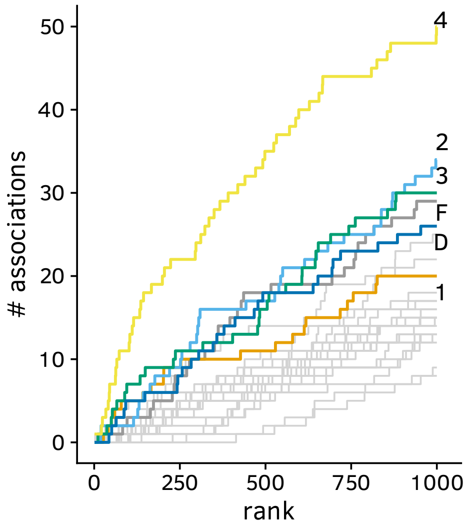
HDV

delta

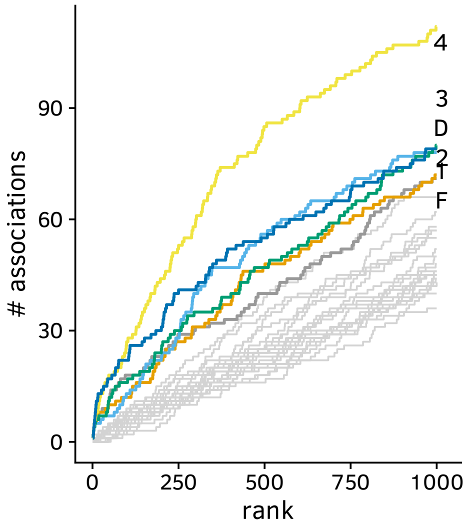


HIV Arevir

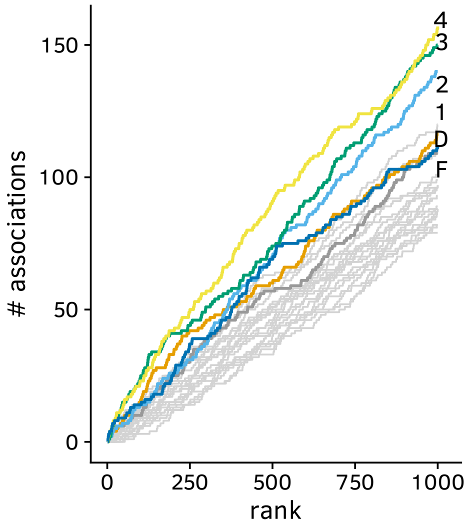
gp120



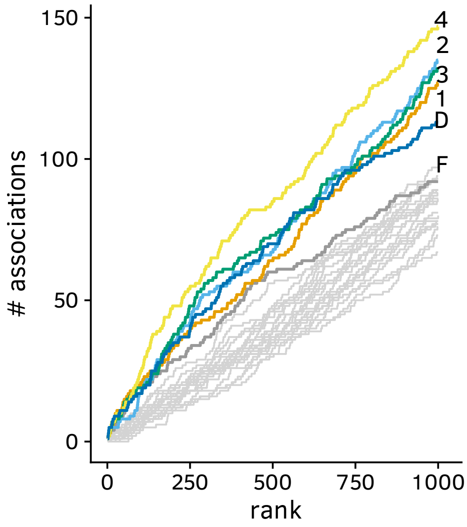
integrase



protease

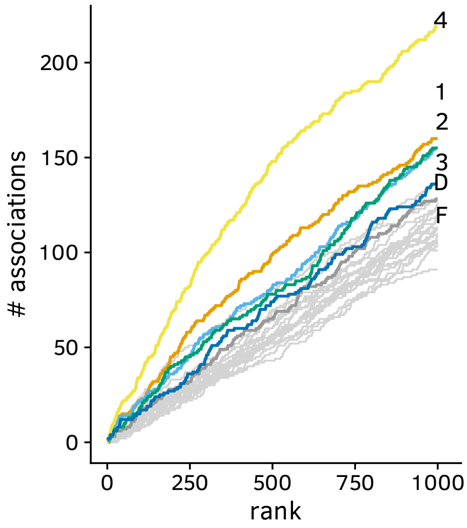


reverse transcriptase

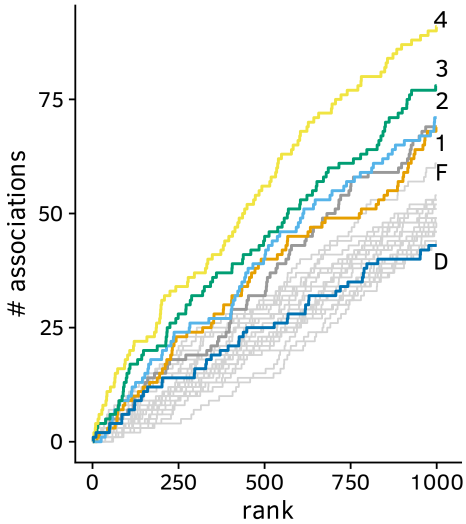


HIV CRF project

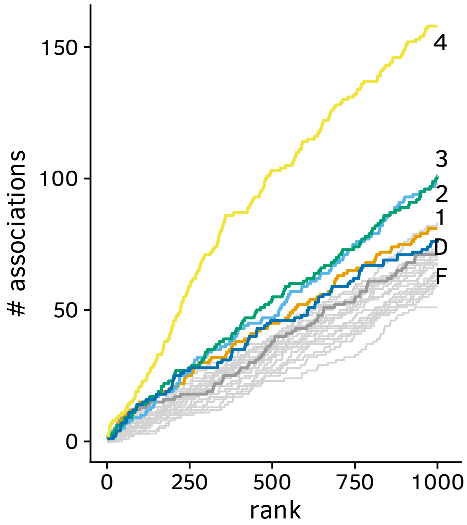
gag



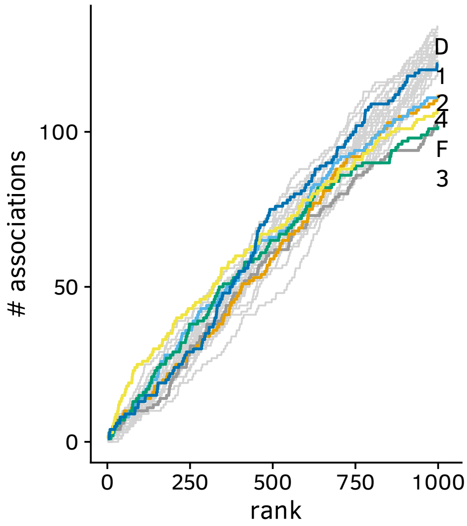
integrase



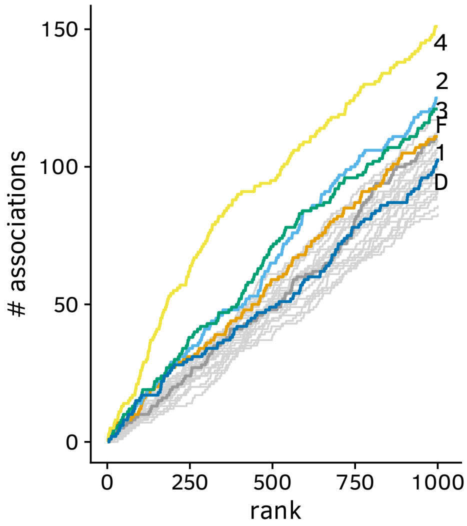
pol



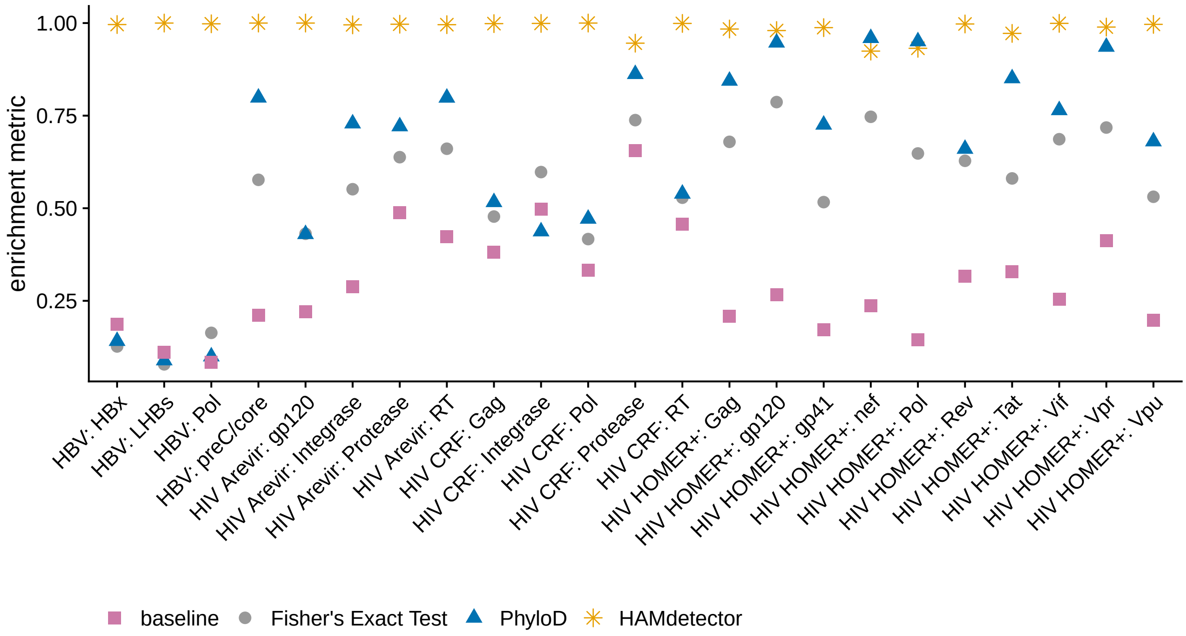
protease



reverse transcriptase



HAM enrichment summary

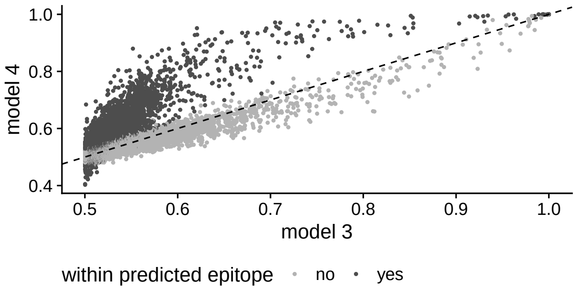


For each model and protein, an enrichment metric is computed according to the following algorithm: We first ranked all evaluated substitutions according to their respective credibility of being a HAM (integral of the marginal posterior $P(\beta_{jk} > 0)$ for HAMdetector, p-values for PhyloD and Fisher's exact testr). Let r be the rank in that sorted list, e.g. rank 10 is the substitution–allele pair at position in that sorted list. Let $N_e(r)$ be the cumulative number of predictions of this rank or better that are located inside known epitopes. Let m_r be the ratio between $N_e(r)$ for model m and $N_e(r)$ for the top performing model. The enrichment metric m is then computed as $m = \frac{1}{N} \sum_{r=1}^{N} m_r$ for ranks 1 to $N = 1000$.

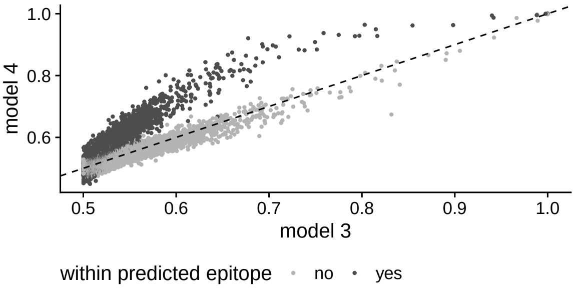
8 Model 3 vs. model 4 scatter plots

HIV HOMER+

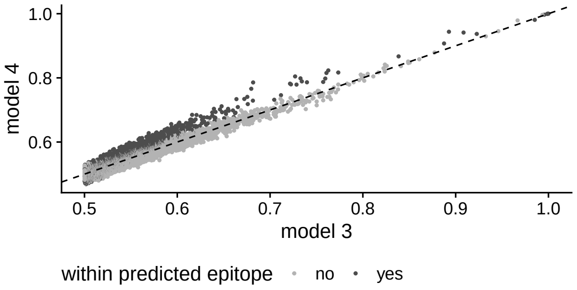
gag



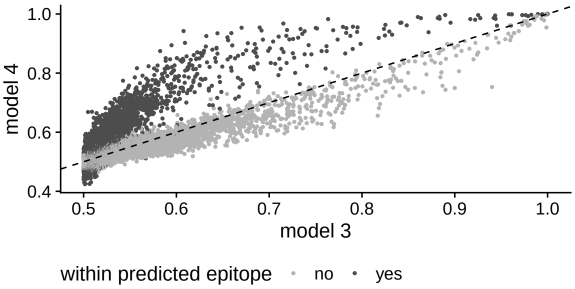
gp41



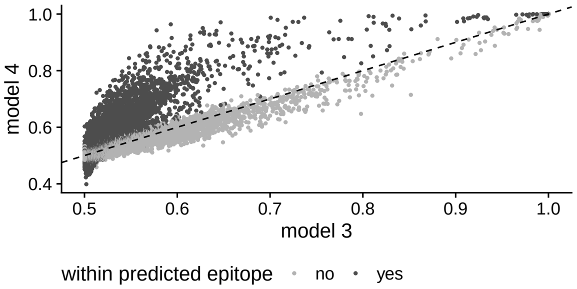
gp120



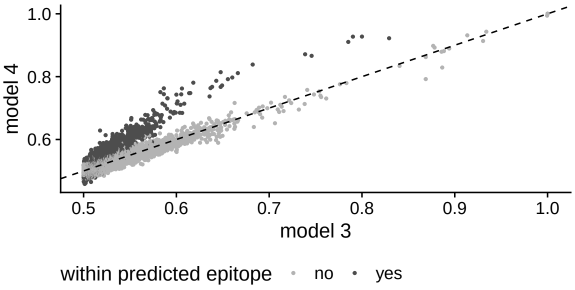
nef



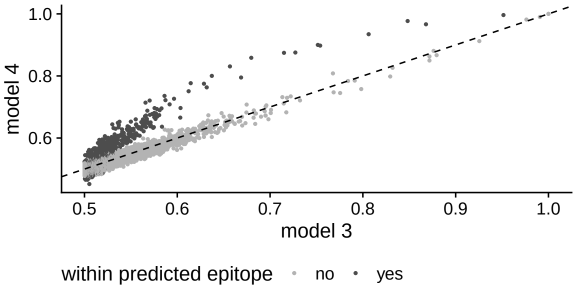
pol



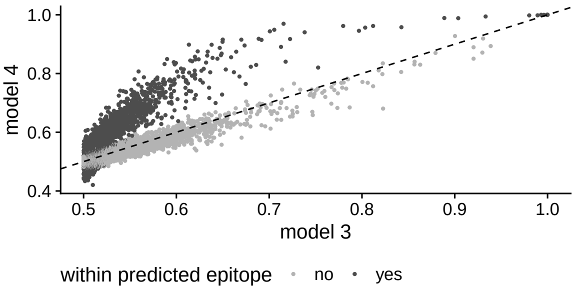
rev



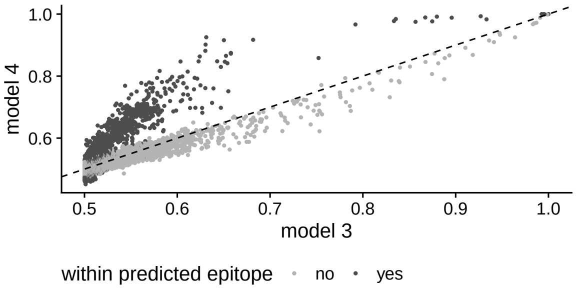
tat



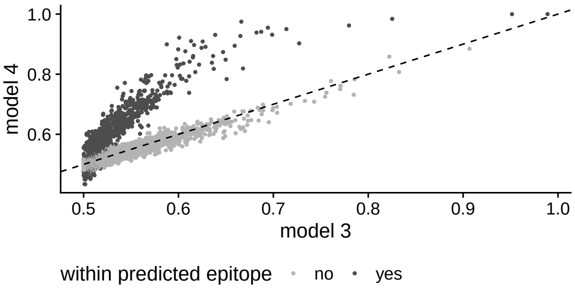
vif



vpr

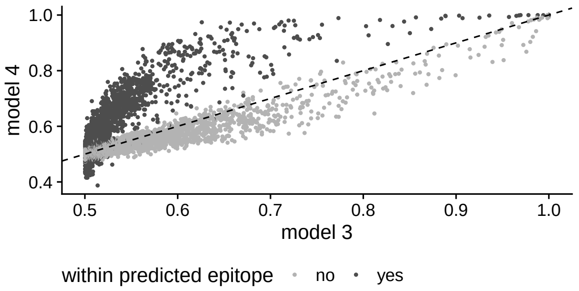


vpu

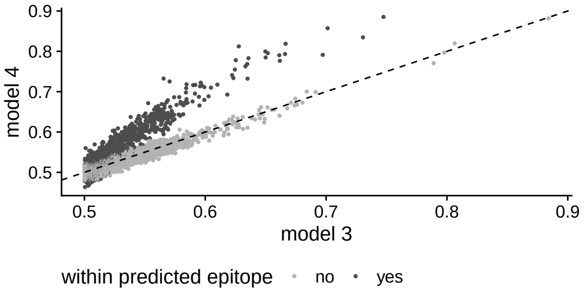


HBV

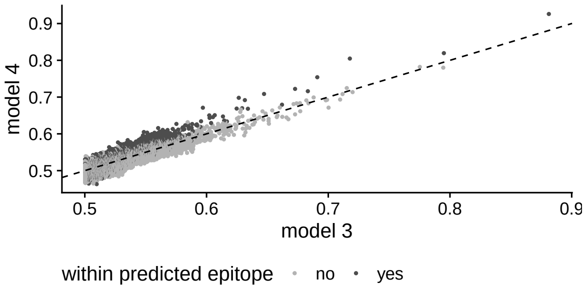
preC/core



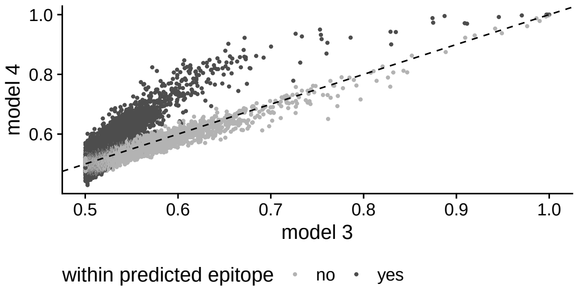
HBx



LHBs

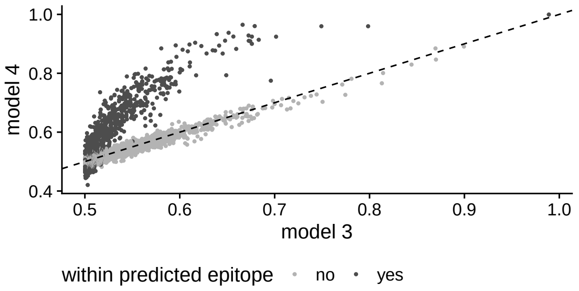


Pol



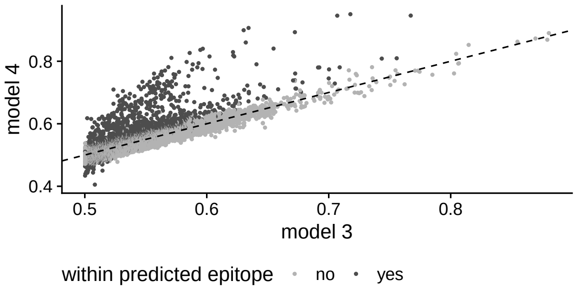
HDV

delta

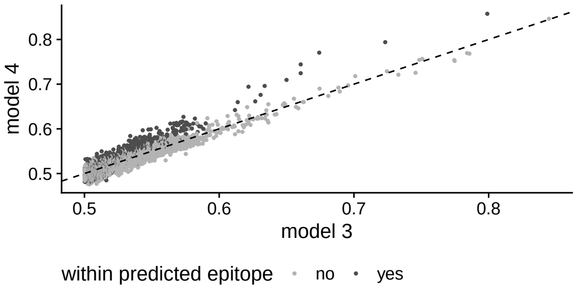


HIV Arevir

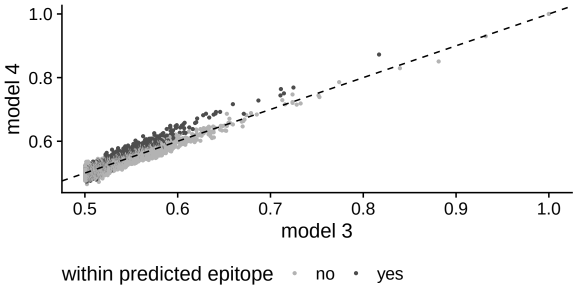
gp120



protease

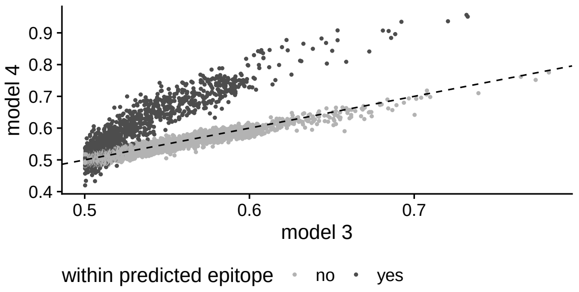


reverse transcriptase

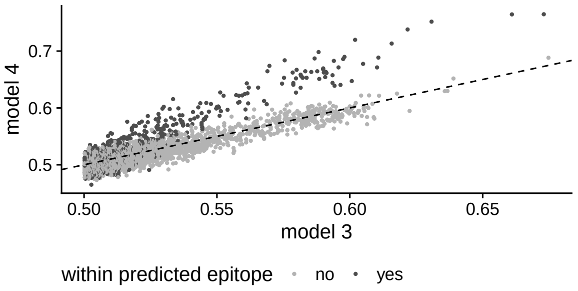


HIV CRF project

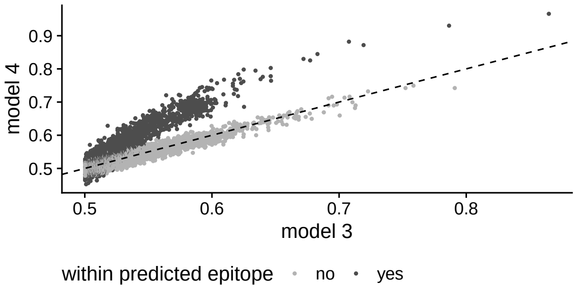
gag



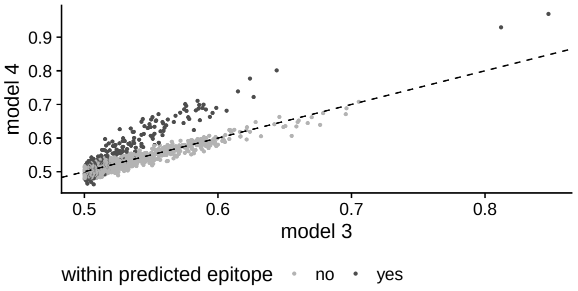
integrase



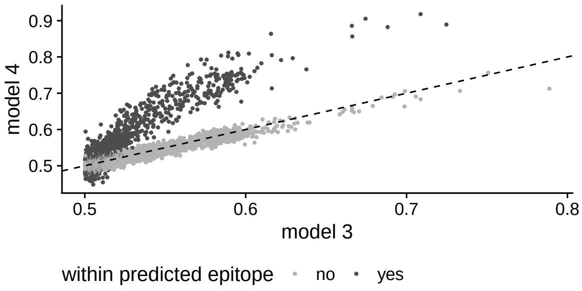
pol



protease



reverse transcriptase



9 LOO comparisons

HIV HOMER+

gag

	elpd_diff	se_diff
logistic regression (baseline)	0.0	0.0
+ horseshoe prior	741.4	46.3
+ phylogeny	3613.5	90.5
+ epitope prediction	13.8	8.1

gp41

	elpd_diff	se_diff
logistic regression (baseline)	0.0	0.0
+ horseshoe prior	7626.7	140.4
+ phylogeny	34068.3	276.5
+ epitope prediction	64.7	16.3

gp120

	elpd_diff	se_diff
logistic regression (baseline)	0.0	0.0
+ horseshoe prior	14068.0	180.1
+ phylogeny	13189.9	172.4
+ epitope prediction	-3.2	20.3

nef

	elpd_diff	se_diff
logistic regression (baseline)	0.0	0.0
+ horseshoe prior	5065.2	119.4
+ phylogeny	28433.4	255.1
+ epitope prediction	51.5	20.1

pol

	elpd_diff	se_diff
logistic regression (baseline)	0.0	0.0
+ horseshoe prior	9627.4	163.4
+ phylogeny	40489.0	321.7
+ epitope prediction	131.8	26.6

rev

	elpd_diff	se_diff
logistic regression (baseline)	0.0	0.0
+ horseshoe prior	3400.4	97.6
+ phylogeny	21225.3	215.9
+ epitope prediction	17.0	9.9

tat

	elpd_diff	se_diff
logistic regression (baseline)	0.0	0.0
+ horseshoe prior	3135.0	91.3
+ phylogeny	19120	205.5
+ epitope prediction	16.5	8.9

vif

	elpd_diff	se_diff
logistic regression (baseline)	0.0	0.0
+ horseshoe prior	4343.0	108.1
+ phylogeny	21254.6	221.3
+ epitope prediction	38.8	16.2

vpr

	$\text{elpd}_{\text{diff}}$	se_{diff}
logistic regression (baseline)	0.0	0.0
+ horseshoe prior	1714.2	71.1
+ phylogeny	10488.2	155.3
+ epitope prediction	23.4	12.4

vpu

	$\text{elpd}_{\text{diff}}$	se_{diff}
logistic regression (baseline)	0.0	0.0
+ horseshoe prior	2877.5	85.4
+ phylogeny	20004.2	199.7
+ epitope prediction	50.1	12.8

HBV

preC/core

	$\text{elpd}_{\text{diff}}$	se_{diff}
logistic regression (baseline)	0.0	0.0
+ horseshoe prior	971.5	65.2
+ phylogeny	4427.8	94.6
+ epitope prediction	65.7	19.5

HBx

	$\text{elpd}_{\text{diff}}$	se_{diff}
logistic regression (baseline)	0.0	0.0
+ horseshoe prior	637.8	52.8
+ phylogeny	8343.2	127.2
+ epitope prediction	8.5	7.9

LHBs

	$\text{elpd}_{\text{diff}}$	se_{diff}
logistic regression (baseline)	0.0	0.0
+ horseshoe prior	1744.2	84.6
+ phylogeny	30120.1	223.8
+ epitope prediction	-7.6	14.7

Pol

	$\text{elpd}_{\text{diff}}$	se_{diff}
logistic regression (baseline)	0.0	0.0
+ horseshoe prior	3157.7	127.7
+ phylogeny	69561.1	343.5
+ epitope prediction	98.8	19.0

HDV

delta

	$\text{elpd}_{\text{diff}}$	se_{diff}
logistic regression (baseline)	0.0	0.0
+ horseshoe prior	455.5	37.5
+ phylogeny	1746.9	63.1
+ epitope prediction	32.1	12.2

HIV Arevir

gp120

	$\text{elpd}_{\text{diff}}$	se_{diff}
logistic regression (baseline)	0.0	0.0
+ horseshoe prior	5453.0	113.1
+ phylogeny	12535.4	167.6
+ epitope prediction	-256.6	33.3

integrase

	$\text{elpd}_{\text{diff}}$	se_{diff}
logistic regression (baseline)	0.0	0.0
+ horseshoe prior	1278.7	58.7
+ phylogeny	6144.4	118.2
+ epitope prediction	1.1	9.9

protease

	$\text{elpd}_{\text{diff}}$	se_{diff}
logistic regression (baseline)	0.0	0.0
+ horseshoe prior	793.7	46.7
+ phylogeny	3883.3	88.3
+ epitope prediction	-19.0	5.7

reverse transcriptase

	$\text{elpd}_{\text{diff}}$	se_{diff}
logistic regression (baseline)	0.0	0.0
+ horseshoe prior	1653.7	65.7
+ phylogeny	7034.2	127.5
+ epitope prediction	-6.1	10.7

HIV CRF project

gag

	$\text{elpd}_{\text{diff}}$	se_{diff}
logistic regression (baseline)	0.0	0.0
+ horseshoe prior	352.2	36.4
+ phylogeny	2893.1	73.1
+ epitope prediction	1.5	13.6

integrase

	$\text{elpd}_{\text{diff}}$	se_{diff}
logistic regression (baseline)	0.0	0.0
+ horseshoe prior	94.4	18.4
+ phylogeny	564.8	36.7
+ epitope prediction	0.0	6.5

pol

	$\text{elpd}_{\text{diff}}$	se_{diff}
logistic regression (baseline)	0.0	0.0
+ horseshoe prior	543.9	42.4
+ phylogeny	3686.2	86.9
+ epitope prediction	-1.9	14.4

protease

	$\text{elpd}_{\text{diff}}$	se_{diff}
logistic regression (baseline)	0.0	0.0
+ horseshoe prior	27.7	13.4
+ phylogeny	297.7	24.5
+ epitope prediction	5.1	4.5

reverse transcriptase

	$\text{elpd}_{\text{diff}}$	se_{diff}
logistic regression (baseline)	0.0	0.0
+ horseshoe prior	198.9	27.3
+ phylogeny	1668.9	56.3
+ epitope prediction	1.7	9.4

10 Software usage

The input for HAMdetector model is aligned viral sequence data that is annotated with the host's HLA type. Optionally, a phylogenetic tree based on the sequence alignment can be used as additional input. An example of (shortened) valid input files is shown below:

`alignment.fasta`

```
>HLA-A32_HLA-A03_HLA-B52_HLA-B14_HLA-C16_HLA-C16
MG-RASVMG-RAV
>HLA-A68_HLA-A31_HLA-B44_HLA-B08_HLA-C07_HLA-C07
MGARASVMG-RTV
>HLA-A24_HLA-A24_HLA-B15_HLA-B46_HLA-C12_HLA-C12
MSARASVMGSRSSV
>HLA-A02_HLA-A03_HLA-B58_HLA-B44_HLA-C14_HLA-C06
MSARASVMG-RRSV
>HLA-A03_HLA-A31_HLA-B55_HLA-B40_HLA-C16_HLA-C03
MSARASVMG-RMSV
```

`phylogeny.tree`

```
((2:0.03,1:0.11):33.61,(3:0.10,5:0.85):0.10,4:0.24);
```

Assuming the files were saved as `/home/user/Desktop/alignment.fasta` and `home/user/Desktop/phylogeny.` HAMdetector can be run using the following commands:

```
using Escape

data = HLAData(
    alignment_file = "/home/user/Desktop/alignment.fasta",
    tree_file = "/home/user/Desktop/phylogeny.tree"
)

result = run_model(data)

summary_df = replacement_summary(result)
```

The output returned by the `replacement_summary` function is a table like the one printed blow:

allele	position	replacement	posterior_p	log_odds_lower_95	log_odds_upper_95
HLAAllele	Int64	Char	Float64	Float64	Float64
HLA-B*55	12	M	0.64375	-1.29526	2.69504
HLA-B*46	12	S	0.6125	-1.43746	2.67338
HLA-B*15	12	S	0.60825	-1.44681	2.54448
HLA-B*40	12	M	0.60175	-1.05725	1.9695
HLA-B*58	12	R	0.59875	-1.62281	2.41285
HLA-A*02	12	R	0.591	-1.55972	2.44342
HLA-C*03	12	M	0.59025	-0.996611	2.03937
HLA-A*31	12	M	0.58875	-1.58969	2.32103
HLA-C*16	12	M	0.5705	-1.62612	2.24687
...	...	...	...	...	...

The table contains a list of all observed replacements, sorted by decreasing estimated probability of being associated with an HLA allele. The column `posterior_p` denotes the posterior probability of the HLA regression coefficient being larger than 0, conditioned on the model and the observed data. Posterior probabilities close to 1 denote a strong positive association between the allele and replacement, posterior probabilities close to 0 denote a strong negative association between the allele and replacement. Biologically, this means: the higher the posterior probability, the stronger the evidence for immune escape.

A posterior probability of 0.5 means that there is no evidence of the replacement being HLA-associated (0.5 because a regression coefficient centered around 0 allocates equal probability to negative and positive values).

3.2 Clinical and molecular characteristics associated with response to therapeutic PD-1/PD-L1 inhibition in advanced Merkel cell carcinoma

This section is based on the following publication:

Ivelina Spassova, Selma Ugurel, Linda Kubat, Lisa Zimmer, Patrick Terheyden, Annalena Mohr, Hannah Björn Andtback, Lisa Villabona, Ulrike Leiter, Thomas Eigentler, Carmen Loquai, Jessica C. Hassel, Thilo Gambichler, Sebastian Haferkamp, Peter Mohr, Claudia Pfoehler, Lucie Heinzerling, Ralf Gutzmer, Jochen S Utikal, Kai Horny, Hans-Ulrich Schildhaus, Daniel Habermann, Daniel Hoffmann, Dirk Schadendorf, Jürgen Christian Becker (2022).
Clinical and molecular characteristics associated with response to therapeutic PD-1/PD-L1 inhibition in advanced Merkel cell carcinoma
`http://dx.doi.org/10.1136/jitc-2021-003198`

Journal for ImmunoTherapy of Cancer

Clinical and molecular characteristics associated with response to therapeutic PD-1/PD-L1 inhibition in advanced Merkel cell carcinoma

Ivelina Spassova ,[1,2] Selma Ugurel ,[2] Linda Kubat,[1,2] Lisa Zimmer,[2] Patrick Terheyden ,[3] Annalena Mohr,[3] Hannah Björn Andtback,[4] Lisa Villabona,[4] Ulrike Leiter,[5] Thomas Eigentler,[5] Carmen Loquai,[6] Jessica C Hassel,[7,8] Thilo Gambichler,[9] Sebastian Haferkamp,[10] Peter Mohr,[11] Claudia Pfoehler,[12] Lucie Heinzerling,[13] Ralf Gutzmer,[14] Jochen S Utikal,[8,15] Kai Horny,[1,8,16] Hans-Ulrich Schildhaus,[17] Daniel Habermann,[18] Daniel Hoffmann,[18] Dirk Schadendorf,[2,16] Jürgen Christian Becker [1,2,8,16]

To cite: Spassova I, Ugurel S, Kubat L, *et al*. Clinical and molecular characteristics associated with response to therapeutic PD-1/PD-L1 inhibition in advanced Merkel cell carcinoma. *Journal for ImmunoTherapy of Cancer* 2022;**10**:e003198. doi:10.1136/jitc-2021-003198

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/jitc-2021-003198).

IS and SU contributed equally.

Accepted 11 December 2021

For numbered affiliations see end of article.

Correspondence to

Professor Jürgen Christian Becker; j.becker@dkfz.de

ABSTRACT

Background Based on its viral-associated or UV-associated carcinogenesis, Merkel cell carcinoma (MCC) is a highly immunogenic skin cancer. Thus, clinically evident MCC occurs either in immuno-compromised patients or based on tumor-intrinsic immune escape mechanisms. This notion may explain that although advanced MCC can be effectively restrained by treatment with PD-1/PD-L1 immune checkpoint inhibitors (ICIs), a considerable percentage of patients does not benefit from ICI therapy. Biomarkers predicting ICI treatment response are currently not available.

Methods The present multicenter retrospective study investigated clinical and molecular characteristics in 114 patients with unresectable MCC at baseline before treatment with ICI for their association with therapy response (best overall response, BOR). In a subset of 21 patients, pretreatment tumor tissue was analyzed for activation, differentiation and spatial distribution of tumor infiltrating lymphocytes (TIL).

Results Of the 114 patients, n=74 (65%) achieved disease control (BOR=complete response/partial response/stable disease) on ICI. A Bayesian cumulative ordinal regression model revealed absence of immunosuppression and a limited number of tumor-involved organ systems was highly associated with a favorable therapy response. Unimpaired overall performance status, high age, normal serum lactate dehydrogenase and normal serum C reactive protein were moderately associated with disease control. While neither tumor Merkel cell polyomavirus nor tumor PD-L1 status showed a correlation with therapy response, treatment with anti-PD-1 antibodies was associated with a higher probability of disease control than treatment with anti-PD-L1 antibodies. Multiplexed immunohistochemistry demonstrated the predominance of CD8$^+$ effector and central memory T cells (T$_{CM}$) in close proximity to tumor cells in patients with a favorable therapy response.

Conclusions Our findings indicate the absence of immunosuppression, a limited number of tumor-affected organs, and a predominance of CD8$^+$ T$_{CM}$ among TIL, as baseline parameters associated with a favorable response to PD-1/PD-L1 ICI therapy of advanced MCC. These factors should be considered when making treatment decisions in MCC patients.

INTRODUCTION

Merkel cell carcinoma (MCC) is a rare, highly aggressive neuroendocrine skin cancer. MCC carcinogenesis is associated either with the Merkel cell polyomavirus (MCPyV), predominantly in cases occurring in the northern hemisphere, or with chronic UV-exposure.[1] MCC is highly immunogenic due to the presence of either MCPyV-derived antigens or UV-associated neoantigens. Thus, clinically manifest advanced MCC is mostly observed in immunocompromised and immunosenescent patients or occurs based on tumor-intrinsic immune escape mechanisms. Still, a high therapeutic activity of PD-1/PD-L1 immune checkpoint inhibitors (ICIs) with durable objective responses in about 50% of patients has been observed.[2] Despite this major improvement in the therapy outcome of advanced MCC patients, this observation also implies that half of the patients do not experience a long-term benefit from ICI therapy. Clinically applicable predictive biomarkers of ICI therapy response are just starting to emerge: (1) in a trial testing neoadjuvant nivolumab, both pathological complete response (CR) and radiographic tumor regression at the time of surgery were correlated with improved recurrence-free survival,[3] and (2) Kacew *et al* reported that a limited disease stage at ICI therapy start was associated with a favorable response.[4]

However, the study also showed that, unlike in other cancers, neither tumor mutational burden nor copy-number alterations in MCC tumor tissue predicted ICI therapy outcome. Similarly, a recent study by us characterizing 41 MCC patients receiving PD-1/PD-L1 ICI demonstrated that predictive markers of ICI therapy response established in other cancer entities such as neutrophil-to-lymphocyte ratio, metastatic stage and site of the primary were not associated with ICI response in MCC.[5] However, our comprehensive dynamic molecular analysis of pretreatment tumor tissue demonstrated that not only the density of the immune cell infiltrate, but rather its functional properties correlated with the response to ICI therapy. In particular, the predominance of central memory T (T_{CM}) cells with a diverse T-cell receptor (TCR) repertoire were associated with a favorable treatment outcome.[5] On the other hand, we did not observe any predictive potential on previously suggested molecular biomarkers such as tumor PD-L1 expression or MCPyV status.[2 5 6] Thus, with the present study we aimed at testing clinically well applicable predictive biomarkers in a larger patient cohort (114 MCC patients, although these included 41 patients from our earlier study[5] by expanding our initial oligocentric approach to a multicentric study, but also limiting the complexity of the molecular analyses to those showing the highest predictive value in our previous study.

MATERIALS AND METHODS

Patients and samples

One hundred and fourteen (n=114) patients treated between May 2018 and July 2020 at 11 MCC referral centers (Bochum, Buxtehude, Erlangen, Essen, Heidelberg, Homburg, Lübeck, Mainz, Regensburg, Stockholm, Tübingen) were retrospectively identified according to the following selection criteria: histopathologically confirmed diagnosis of MCC, treatment with PD-1/PD-L1 ICI for unresectable advanced disease, and complete follow-up documentation of ICI therapy outcome including best overall response (BOR), progression-free (PFS) and overall (OS) survival. BOR was determined according to RECIST V.1.1.[7] PFS and OS were defined as time from therapy start until disease progression or death, respectively; if no such event occurred, the date of the last patient contact was used as endpoint of survival assessment (censored observation). Detailed clinical parameters at baseline of ICI therapy were collected from the patients' medical charts; it is important to note that a subgroup of 41 patients had already been described in an earlier study[5] (online supplemental table S1). Immunosuppression was assigned to patients suffering from hematological neoplasia or to patients treated with multiple drugs for multiple cancers or immunosuppressive medications. If available, pretreatment samples of formalin-fixed paraffin-embedded (FFPE) tumor tissue from the studied patients were collected for molecular analysis.

Detection of MCPyV DNA

Detection of MCPyV DNA was performed as previously described by TaqMan Real-Time qPCR using the following large T-antigen (LTA) specific primers and TaqMan probe: forward primer; CCA AAC CAA AGA ATA AAG CAC TGA; reverse primer, TCG CCA GCA TTG TAG TCT AAA AAC, and probe: FAM-AGC AAA AAC ACT CTC CCC ACG TCA GAC AG-BHQ1.[5]

Multiplex immunofluorescence staining

Multiplex immunofluorescence staining of FFPE tumor tissue was performed using the Opal chemistry (PerkinElmer, Waltham, USA, Cat.No.: OP7TL4001KT) with two panels of antibodies, ie, against CD4, CD8, CD20, Foxp3 and CD68 (panel 1), or CD27, GZMB, TCF1, CD45RA and CD45RO (panel 2). Synaptophysin served as tumor marker in either panels. Briefly, after deparaffinization and fixation, 3 µm tumor sections were processed with retrieval buffers for 15 min in an inverter microwave oven. Thereafter, sections were incubated with the antibody diluent for 10 min at room temperature, followed by incubation with the primary antibody for 30 min. After applying Opal polymer HRP secondary antibody and Opal fluorophore solution each for 10 min, antibodies were removed by microwave treatment before a further round of staining. The antibodies, their dilutions, the respective retrieval buffers as well as the sequence of usage are described in detail in online supplemental table S2. Visualization of the different fluorophores was achieved on the Mantra Quantitative Pathology Imaging System (PerkinElmer). For each tumor sample, quantification of the different cell types was performed at medium magnification on three randomly selected areas located either in the juxtatumoral or intratumoral region in a semiautomatic fashion with the InForm Tissue Analysis software (Akoya Biosciences, Menlo Park, USA). Since tumor samples were received from different pathology institutes, the quality of the FFPE material was not uniform, resulting in variations in fluorescence intensity from sample to sample. To avoid quantification errors due to these intensity variations, the InForm tissue analysis software was trained on tumor tissue samples from the respective sources, thus developing an algorithm based on the median of the determined intensities. Subsequently, training of the software was performed on five different MCC tissue samples to recognize staining patterns/cell types. Finally, a principal components analysis was used to visualize possible pattern of the immunofluorescence staining results across the samples as suggested by Shen *et al*.[8] Annotation by the different pathology institutes the samples were received from no significant batch effect was observed indicating the initial training of the InForm software was sufficient.

Two independent observer, blinded to ICI response, monitored the quantification analysis and classified the respective cell types in relation to all nucleated cells per sample into five categories: 0%, >1%, 1%–5%, 5%–10% and >10%. Disagreements were resolved by taking the opinion of a third observer into consideration. Markers used for the quantification of the different immune cell types are listed in online supplemental table S3 and the raw data of the quantification analysis by InForm Tissue Analysis software is provided in online supplemental table S4 and S5 for panel 1 and panel 2, respectively.

STATISTICAL ANALYSES

A Bayesian cumulative ordinal regression model was applied for predicting PD-1/PD-L1 ICI therapy response in MCC patients. The model was fit with a dataset consisting of 114 patients and 15 clinical parameters; year of treatment and participation in a clinical trial were considered as possible confounders (table 1). Model parameters were described with probability distributions that take into account the uncertainty of the estimates. Treatment response was classified into CR, partial response (PR), stable disease (SD) and progressive disease (PD) on an ordinal scale. We applied a cumulative ordinal regression model to the data, which takes the ordinal nature of the response variable into account. Compared with other approaches that incorrectly treat the response variable as metric (such as linear regression) or nominal (eg, by binarizing the response variable), this may lead to more precise inference and therefore reduces over- or underestimation of effect sizes.[9] The cumulative regression model regards the tendency of a patient to respond to treatment as a latent (=unobserved) variable that is determined by the patient characteristics. The model is described by the following formula: $_i = \beta_{age}.x_{i,age} + \beta_{LDH}.x_{i,LDH} \dots$; in which is the location of the latent variable for patient i, β_{age} is the β coefficient for the predictor age and $x_{i,age}$ is the indicator variable of patient i for age. The β coefficient of the predictor provides information on whether or not a predictor is associated with a higher probability of treatment response. Student *t*-priors with 7 degrees of freedom and a SD of 1 were chosen as weakly informative priors for the β coefficients. Model fit is performed numerically by Markov chain Monte Carlo.[10] The width of the distribution gives an impression of the uncertainty of the estimate: a distribution tightly concentrated around a value means that the dataset allows for a precise estimate of that parameter, while a broader distribution means that the data is consistent with a wide range of parameter values. Average predictive comparisons are calculated as expected changes in response associated with a unit difference in one of the inputs. They were calculated with respect to having at least a PR to treatment, for example, for immunosuppression, values between −0% and −40% denote that comparing a patient with immunosuppression to an otherwise identical patient without immunosuppression, the patient with immunosuppression has (on average) a 0%–40% lower probability of having at least a PR to treatment. Fitting the model to the dataset was done with the R software package 'brms', which utilizes 'Stan' in the background.[11] Missing values were estimated by multiple imputation with the R package 'mice'.[11–13] The imputed data values are consistent with the observed data (online supplemental figure 1). Using leave-one-out cross-validation (LOO), this model has a similar (expected log-predictive density (ELPD), a measure of its ability to generalize to unseen data) as a sequential model without category-specific effects, meaning that including category-specific effects does not improve model performance and the proportional odds assumption does not have a strong effect on the model conclusions. In the following, we report the detailed results for LOO and ELPD for completeness (online supplemental figure 2). A detailed description is given in online supplemental materials and methods. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis reporting guidelines were followed to develop the predictive model, including patient selection.[14]

Kaplan-Meier plots were generated with R V.3.5.1 using the package 'survival' (V.2.44–1.1 and survminer V.0.4.6). All patients with PFS and/or OS of more than 36 months are censored without having a respective event, because the data beyond this period is very sparse. Log rank test is used to calculate p values.

For statistical testing of T-cell abundance in MCC tumor tissue, p values were determined by beta regression with R V.4.0.2 and packages 'lmtest' and 'betareg' (V.0.9–38 and V.3.1–4).

For statistical testing of the distance between tumor cells and CD8$^+$ T cells in tumor tissue, the unpaired two-tailed Student's t-test, calculated in GraphPad Prism V.5 (San Diego, USA) was applied.

RESULTS

Patient characteristics, response to ICI and survival outcomes

A total of 114 patients treated with PD-1/PD-L1 ICI (avelumab, n=57; nivolumab, n=13; pembrolizumab, n=44) for unresectable advanced MCC were identified at 11 MCC referral centers in Germany and Sweden. Detailed patient characteristics are given in table 1. Of 114 patients, 54 (47%) experienced an objective response (BOR=CR/PR), and 74/114 patients (65%) a disease control (BOR=CR/PR/SD) on ICI (figure 1A). When the entire patient cohort was divided by type of therapy, in the cohort treated with the anti-PD-L1 antibody avelumab, 22/57 patients (39%) experienced an objective response and 33/57 patients (58%) disease control; in the cohort treated with the anti-PD-1 antibodies pembrolizumab or nivolumab, an objective response was observed in 32/57 patients (56%) and disease control in 41/57 patients (72%). Of 114 patients, 37 (32%) died within a median follow-up time of 12.0 (±2.41, 95% CI) months. Kaplan-Meier estimates for PFS and OS categorized by BOR (CR, n=24; PR, n=30; SD, n=20; and PD, n=40) revealed a clear separation of the curves for patients experiencing disease control (CR/PR/SD) as BOR compared with those presenting a primary progression on therapy (online supplemental figure 3). The median PFS and OS in the

Table 1 Patient and tumor characteristics at baseline of anti-PD-1/PD-L1 therapy

	All patients n=114 (100%)	Disease control (BOR=CR/PR/SD) n=74 (100%)	Disease progression (BOR=PD) n=40 (100%)
Patient characteristics			
Gender			
Male	82 (72%)	54 (73%)	28 (70%)
Female	32 (28%)	20 (27%)	12 (30%)
Age			
<70	40 (35%)	24 (32%)	16 (40%)
≥70	74 (65%)	50 (68%)	24 (60%)
Overall performance status (ECOG)			
0	64 (56%)	47 (64%)	17 (43%)
≥1	49 (43%)	26 (35%)	23 (57)
Not available	1 (1%)	1 (1%)	0 (0%)
Immunosuppression			
No	92 (81%)	64 (86%)	28 (70%)
Yes	22 (19%)	10 (14%)	12 (30%)
LDH (blood)			
Normal	43 (38%)	32 (43%)	11 (28%)
Elevated	67 (59%)	39 (53%)	28 (70%)
Not available	4 (3%)	3 (4%)	1 (2%)
CRP (blood)			
Normal	30 (26%)	21 (28%)	9 (23%)
Elevated	55 (48%)	31 (42%)	24 (60%)
Not available	29 (26%)	22 (30%)	7 (16%)
NLR (blood)			
<4	54 (47%)	36 (49%)	18 (45%)
≥4	35 (31%)	20 (27%)	15 (38%)
Not available	25 (22%)	18 (24%)	7 (17%)
Tumor characteristics			
Localization of primary			
Head and neck	24 (21%)	17 (23%)	7 (17%)
Extremities	44 (39%)	27 (36%)	17 (43%)
Trunk	19 (17%)	12 (16%)	7 (17%)
Unknown primary	15 (13%)	9 (12%)	6 (15%)
Metastatic stage (AJCC)			
M0	17 (15%)	11 (15%)	6 (15%)
M1a	36 (32%)	23 (31%)	13 (32%)
M1b/M1c	61 (53%)	40 (54%)	21 (53%)
Organs involved			
1	51 (45%)	38 (51%)	13 (32%)
>1	63 (55%)	36 (49%)	27 (68%)
MCPyV status (tumor)			
Negative	10 (9%)	6 (8%)	4 (10%)
Positive	32 (28%)	20 (27%)	12 (30%)
Not available	72 (63%)	48 (65%)	24 (60%)
PD-L1 (tumor)			
Negative	17 (15%)	11 (15%)	6 (15%)

Continued

Table 1 Continued

	All patients n=114 (100%)	Disease control (BOR=CR/PR/SD) n=74 (100%)	Disease progression (BOR=PD) n=40 (100%)
Positive	21 (18%)	12 (16%)	9 (23%)
Not available	76 (67%)	51 (69%)	25 (62%)
Therapeutic interventions			
Previous radiotherapy			
No	55 (48%)	35 (47%)	20 (50%)
Yes	59 (52%)	39 (53%)	20 (50%)
Previous chemotherapy			
No	83 (73%)	53 (72%)	30 (75%)
Yes	31 (27%)	21 (28%)	10 (25%)
PD-1/PD-L1 inhibitor therapy			
Avelumab	57 (50%)	33 (45%)	24 (60%)
Nivolumab	13 (11%)	10 (13%)	3 (8%)
Pembrolizumab	44 (39%)	31 (42%)	13 (32%)

AJCC, American Joint Committee on Cancer; BOR, best overall response; CRP, C reactive protein; ECOG, Eastern Cooperative Oncology Group; LDH, lactate dehydrogenase; MCPyV, Merkel cell polyomavirus; NLR, neutrophil to lymphocyte ratio.

control group were 12.1 and 15.9 months, and 1.4 and 3.9 months, respectively, in the progression group.

On the finding of this clear separation in survival probabilities between patients responding with disease control (BOR=CR/PR/SD) and patients responding with disease progression (BOR=PD), we performed all further molecular analyses on the association of clinical and molecular characteristics with therapy response based on the discrimination between these two patient groups (ie, disease control group vs disease progression group). Due to the limited number of samples, we refrained from forming further subclusters taking the degree of response into account.

Baseline clinical parameters are associated with a favorable response to ICI therapy

Most predictive models dichotomize response to therapy, which neglects the extent of the response. To overcome this limitation, we developed a Bayesian model that instead of dichotomizing the therapy response into two groups, *that* is, regression *versus* progression, rather reflects the established clinical response evaluation

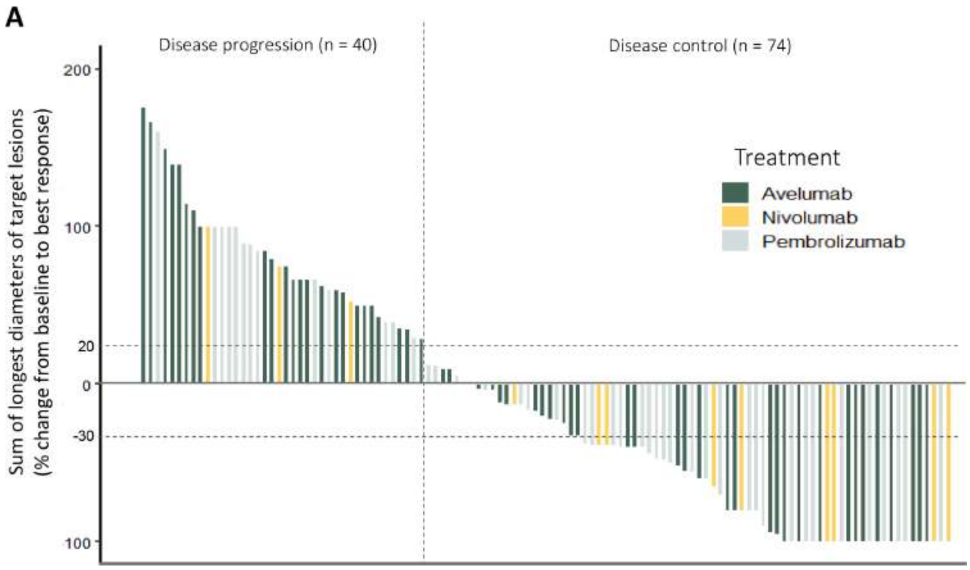


Figure 1 Response of n=114 advanced MCC patients on PD-1/PD-L1 immune checkpoint inhibition therapy. Waterfall plot depicting the best overall response (BOR) as change in the sum of the longest diameters of target lesions from baseline to BOR. Each bar, color coded by therapeutic antibody, represents an individual patient. The pointed vertical line discriminates patients with disease control (BOR=CR/PR/SD) from patients with disease progression (BOR=PD). CR, complete response; MCC, Merkel cell carcinoma; PD, progressive disease; PR, partial response; SD, stable disease.

criteria in solid tumors (RECIST): CR, PR, SD, and PD.[7] Resulting from this Bayesian model, we found the absence of immunosuppression as well as a limited number of organs (1 vs>1) involved into disease spread as the strongest predictors of a favorable response to PD-1/PD-L1 ICI therapy (figure 2). Interestingly, the involved organ type, for example, soft tissue versus visceral, showed less predictive power. These calculations indicate that immunocompetent patients or patients with only one affected organ have probability to achieve disease control on ICI treatment, that is by about 20% higher (0%–40% increase contains almost all the probability). Additionally, an unimpaired overall performance status (Eastern Cooperative Oncology Group (ECOG)=0), patient age of 70 years and above, as well as normal lactate dehydrogenase (LDH) and C reactive protein (CRP) serum levels were associated with a higher probability of disease control on ICI, but to a lower extent. Interestingly, patients' sex, localization of the primary, pretreatment with radiation or chemotherapy, and MCPyV status or PD-L1 expression of the tumor revealed no relevant association with ICI therapy response. Similarly, when we tested if year of treatment and participation in a clinical trial were possible confounders, no impact on the therapeutic outcome was observed. Surprisingly, in our investigated patient cohort with an equal distribution of PD-L1 and PD-1 ICI therapies (see table 1), the use of anti-PD-1 antibodies for ICI therapy was associated with a higher probability of a favorable therapy response (figure 2). Notably, the distribution of relevant patient and tumor characteristics, particularly immunosuppressive state, number of organs involved with disease, impaired ECOG status, and elevated serum LDH and CRP were equally distributed among the two treatment cohorts, that is, anti-PD-1 and anti-PD-L1 antibody ICI (online supplemental table S6).

Pretreatment dense intratumoral infiltrates of CD8$^+$ T$_{CM}$ are associated with a favorable response to ICI

We recently demonstrated by transcriptomics, spatial proteomics and TCR sequencing of sequential tumor biopsies under PD-1/PD-L1 ICI therapy that a predominance of T$_{CM}$ with a diverse TCR repertoire and the ability to expand on ICI is associated with a favorable therapy response.[5] This approach allows a good understanding of the complex immune biology of MCC, but is difficult to use in the clinical routine of patient care. In order to establish clinically well applicable predictive biomarkers, we here limited the complexity of our molecular investigations to multiplexed immunohistochemistry of pretreatment FFPE tumor tissue in order to extract the most important cell type characteristics for therapy response, which can be realistically analyzed as predictive biomarker in the future. Phenotyping of the immune infiltrate of pretreatment tumor tissue samples of 21 patients, including some tumors from our earlier report,[5] (11 patient with disease control, 10 patients with disease progression) for the expression of CD4, CD8, CD20, Foxp3, and CD68 showed that dense immune cell infiltration, particularly by CD8$^+$ T cells, correlated with a favorable ICI therapy response (figure 3A,B). Significantly more CD8$^+$ T cells were infiltrating the juxtatumoral area (p=0.02) and higher number of cytotoxic T cells were present in the intratumoral area (p=0.16) of patients achieving disease control (figure 3C, online supplemental figure 4). Figure 3D illustrates how the spatial distribution of the tumor-infiltrating CD8$^+$ T cells was measured as the distance between the nuclei of CD8$^+$ T cells and synaptophysin$^+$ MCC cells. For distance analysis, tumor samples with CD8$^+$ T cells that were less than 1% of total cells had to be excluded, because in these cases it was not possible to measure the distance of at least 20 different tumor/T–cell pairs. In patients with disease control, CD8$^+$ T cells were in direct and close contact with the tumor cells with a mean distance length of 13.24 µm, whereas the mean distance length was significantly higher in patients with disease progression (22.00 µm, p=0.009) (figure 3E). Moreover, in patients showing disease progression, the CD8$^+$ T cells were mostly restricted to the juxtatumoral stromal space, and only rarely within the tumor tissue. We did not detect significant differences in the amount or distribution of regulatory T cells (CD4$^+$FoxP3$^+$), B cells (CD20$^+$), and monocytes/macrophages (CD68$^+$) between tumor tissues of patients with disease control and patients with disease progression. However, with respect to CD20$^+$ B cells within the cellular immune infiltrate, it is important to note that their frequency varied strongly between samples.

Staining for CD27, GZMB, TCF1, CD45RA and CD45RO allows a precise distinction of T$_{CM}$ and effector T cells with T$_{CM}$ characterized by colocalization of CD27, TCF1 and CD45RO. Indeed, only in pretreatment tumors from patients with disease control, we observed a clear co-localization of these T$_{CM}$ markers (figure 4). Moreover, quantification of T-cell subtypes confirmed a higher percentage of T$_{CM}$ of total tumor infiltrating lymphocytes (TILs) number in the intratumoral infiltrate as well as in the juxtatumoral area (table 2). Effector T cells characterized by colocalization of GZMB and CD45RA were also more frequently observed in the cellular tumor infiltrate of pretreatment tumor tissue of patients with disease control than in those with disease progression; however, the difference was less evident (p=0.07; online supplemental figure 5). It should be noted that CD45RA re-expression has also been described in terminally differentiated T cells characterized by decreased proliferative capacity, increased senescence signaling in vitro.[15]

DISCUSSION

To find predictors, which assess the individual probability of success of PD-1/PD-L1 ICI in MCC, we collected clinical information on 114 accordingly treated patients and established the spatial distribution of tumor infiltrating T cells as well as their activation and differentiation status in pretreatment FFPE tumor tissue samples by two IHC panels. We developed a Bayesian cumulative ordinal

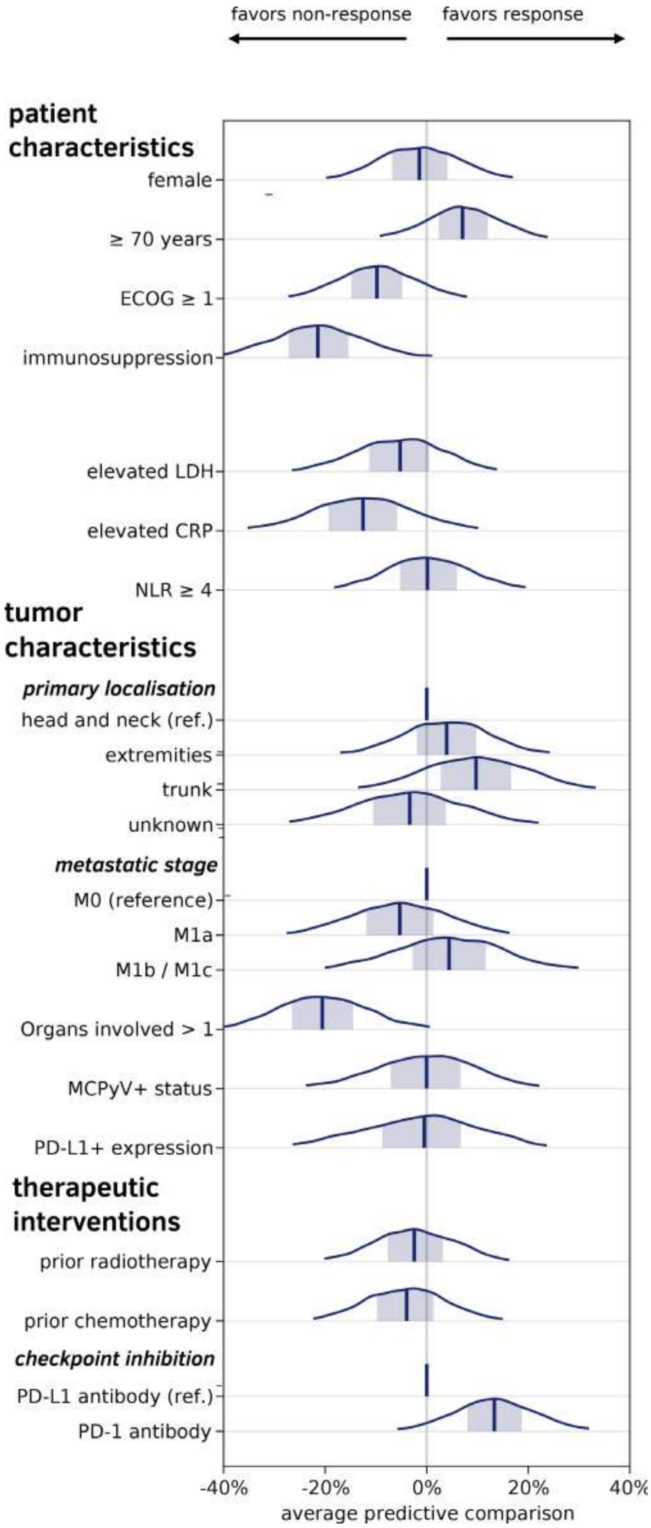


Figure 2 Best overall response (BOR) to anti-PD-1/PD-L1 therapy in correlation to baseline clinical patient and tumor characteristics. The correlations are visualized by average predictive comparisons calculated by a Bayesian cumulative ordinal regression model. While the presented data refer to the full model using four categories of response: CR, PR, SD, and PD, to ease interpretation we mapped the obtained results by average predictive comparisons on a single probability scale for disease control (BOR=CR/PR/SD) and disease progression (BOR=PD) as a probability distribution, given as the percentage of average predictive comparison. The 95% credibility intervals are colored in light blue. Distinct parameters are marked as reference (Ref), described as vertical blue lines set at 0% average predictive comparison. CR, complete response; CRP, C reactive protein; ECOG, Eastern Cooperative Oncology Group; LDH, lactate dehydrogenase; MCPyV, Merkel cell polyomavirus; NLR, neutrophil to lymphocyte ratio; PD, progressive disease; PR, partial response; SD, stable disease.

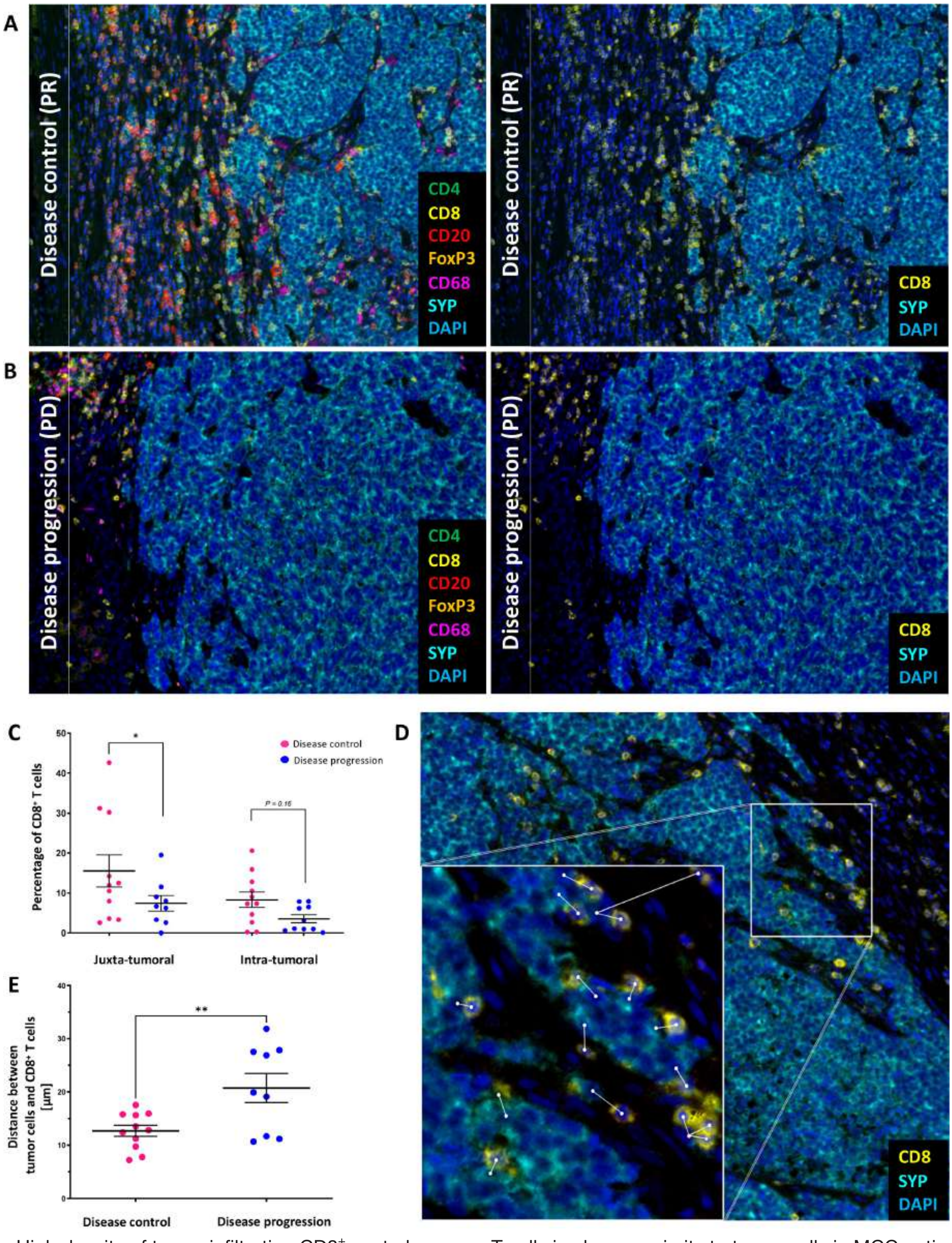


Figure 3 High density of tumor-infiltrating CD8$^+$ central memory T cells in close proximity to tumor cells in MCC patients showing disease control (CR/PR/SD) on PD-1/PD-L1 ICI phenotyping of the cellular immune infiltrate present in MCC tumor lesions obtained at baseline of ICI therapy of a representative patient responding with disease control (A) and disease progression (B) was done by multiplexed immunohistochemistry-based staining using antibodies against CD4 (green), CD8 (yellow), CD20 (red), FOXP3 (orange), CD68 (magenta), and the MCC marker synaptophysin (SYN) (cyan); nuclei are stained with DAPI (blue). depicted are merged images at ×20 magnification. (C) Percentage of CD8$^+$ T cells in pretreatment tumor tissue from patients showing disease control and those showing disease progression in the juxtatumoral and intratumoral area. P values were determined using beta regression. (D) Measurement of the distance between CD8$^+$ T cells and tumor cells. (E) Mean value of the distance between CD8$^+$ T cells and tumor cells for patients showing disease control and those showing disease progression. P values were determined using unpaired, two-tailed Student's t-test. CR, complete response; ICI, immune checkpoint inhibitor; MCC, Merkel cell carcinoma; PR, partial response; SD, stable disease.

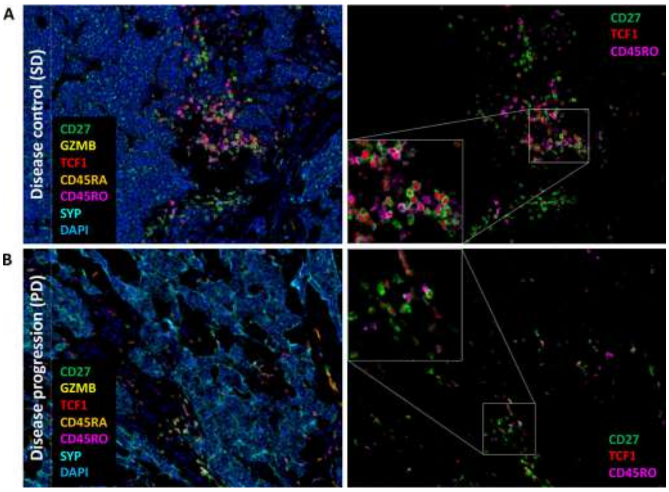


Figure 4 Predominance of central memory T cells (T_{CM}) among tumor-infiltrating lymphocytes of patients showing disease control (CR/PR/SD) on PD-1/PD-L1 ICI therapy. Multiplexed immunofluorescence staining of pretreatment tumor tissue from a representative patient showing disease control (A) and disease progression (B) using antibodies against CD27 (green), GZMB (yellow), TCF1 (red), CD45RA (orange), CD45RO (magenta), and the MCC marker synaptophysin (SYN) (cyan); nuclei are stained with DAPI (blue). Depicted are merged images at ×20 magnification. To visualize the colocalization of CD27, TCF1 and CD45RO, an enlarged image view is shown. CR, complete response; ICI, immune checkpoint inhibitor; MCC, Merkel cell carcinoma; PR, partial response; SD, stable disease.

regression model that includes the distance between clinical characteristics and thereby appropriately accounts for category order. This model avoids problems such as dichotomizing the outcome or treating the distance between the categories as equal, and thus uses the available data efficiently. It revealed the absence of immunosuppression and the metastatic involvement of a limited number of organ systems as characteristics predicting disease control on PD-1/PD-L1 ICI therapy with the highest probability. Additional characteristics associated with treatment response were age, overall performance status, serum LDH and serum CRP, as well as a brisk intratumoral infiltrate by T_{CM} (figure 5). However, both the data model and the missing data model rely on assumptions about the data generating process, for example, that data are missing at random. Even if the implications of these assumptions have been evaluated carefully, all results are still conditioned on the underlying model and should be interpreted with this in mind.

The expanded patient cohort under investigation of n=114 that allowed fitting a more complex Bayesian model supported our previously published observations in 41 patients.[5] The positive effect of an intact immune system observed in either study was also reported from the avelumab expanded access program for metastatic MCC patients in which immunocompromised patients achieved a lower response rate with shorter durations of response.[16] It should be noted that other characteristics associated with a lower probability of response, such as a limited performance status (ECOG >0) or an elevated serum CRP, are also likely to reflect an impaired immune status of the respective patient. Similarly, an elevated serum LDH level and a higher number of involved organs are not to be interpreted only as markers of a higher tumor load. Notably, in our previous report, we dichotomized the number of involved organs up to two or more, these groups did not show a clear association with the probability of response but had broad posterior intervals; in contrast, the larger cohort is consistent with the involvement of only one organ being a strong positive predictive marker. Patient age of 70 years and above was also found to be associated with a higher probability of disease control, but to a lesser extent. A positive correlation of response to anti-PD-1 therapy and patient age ≥60 years was described in melanoma.[17] In this respect, it is important to note that chronological age does not necessarily reflect immunological age. One factor that correlates better with biological/immunological age than chronological age is frailty, which directly describes a person's health status.[18] For example, the process of inflammation is a predictor for frailty and one of the key cell types believed to facilitate an inflammatory phenotype are tumor-infiltrating macrophages, which are often detected in MCC tissue.[19] Other clinical characteristics such as sex, primary site, prior radiation or chemotherapy, PD-L1 expression by tumor cells, and tumor MCPyV status did not show relevant association to PD-1/PD-L1 ICI response. These observations are in line with the results from a study scrutinizing 37 MCC patients receiving ICI.[20]

Functional characterization of the immunological infiltrate of pretreatment tumor tissue revealed that in particular the presence of T_{CM} in close proximity to tumor cells was associated with a favorable response to ICI therapy. It is important to note that because the tumor samples were received from different pathology institutes, the quality of the FFPE material was not uniform, resulting in variations in fluorescence intensity from sample to sample. To avoid quantification errors due to these intensity variations, the InForm tissue analysis software was trained on tumor tissue samples from the respective sources, thus developing an algorithm based on the median of the determined intensities. Moreover, the observation was consistent with our previous work, where we performed transcriptomics, spatial proteomics and TCR sequencing of sequential tumor biopsies before and under ICI therapy of rather uniform quality. These observations confirm the robustness of the chosen approach and the importance of T_{CM} as one of the effectors of response to ICI therapy. The superior antitumor efficacy of T_{CM} cells can be explained by their low activation threshold, rapid proliferative and differentiation capacity on cognate activation, as well as their capacity for long-term persistence facilitating immunologic memory.[21 22] Indeed, since T_{CM} cells are the major source of secondary effector cells during a recall response, the duration of anti-tumor immune responses depends

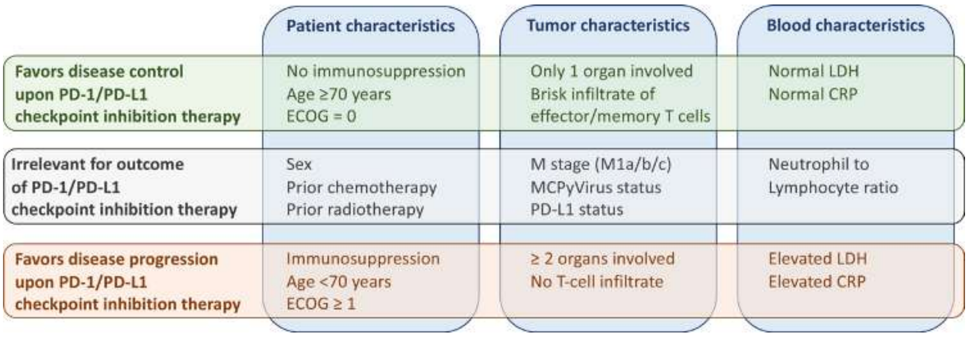
Table 2 Quantification of the cellular tumor infiltrate characterized by multiplex immunohistochemistry staining

Pretreatment MCC tumor tissue samples

		Total leucocyte no per observed area		T_{CM} in % of total lymphocyte no per observed area	
Response to CPI		**Juxtatumoral**	**Intratumoral**	**Juxtatumoral**	**Intratumoral**
Disease control	PR	2121	731	4.7	2.2
	CR	1510	989	3.0	2.7
	CR	1813	3519	0.0	22.0
	SD	5436	9405	18.0	11.7
	PR	1902	2162	6.7	8.0
	PR	1451	97	13.0	26.7
	SD	115	283	23.0	33.0
	CR	1	14	0.0	0.0
	PR	1846	6030	0.0	11.0
	PR	1681	2080	13.0	11.7
	CR	2261	428	33.0	36.7
	Mean value	1831	2340	10.4	15.1
Disease progression	PD	*NA*	0	*NA*	0.0
	PD	44	54	0.0	0.0
	PD	1125	124	0.0	0.0
	PD	239	170	12.0	15.0
	PD	3831	2478	22.0	13.0
	PD	316	376	17.0	10.0
	PD	0	13	0.0	0.0
	PD	2438	287	3.7	0.5
	PD	403	62	2.0	1.7
	PD	159	66	0.0	0.0
	Mean value	950	363	6.3	4.0

Tumor tissue samples were obtained from MCC patients prior to the start of PD-1/PD-L1 immune checkpoint inhibitor therapy. Lymphocytes were identified based on $CD45RA^+$ or $CD45RO^+$ staining and the sum of both signals were used for the quantification of the total lymphocyte number per sample per observed area. T_{CM} were determined based on the triple $CD27^+TCF1^+CD45RO^+$ staining.
CPI, checkpoint inhibition; CR, complete response; MCC, Merkel cell carcinoma; PD, progressive disease; PR, partial response; SD, stable disease; T_{CM}, central memory T-cells.

on their presence.[23] This is consistent with our present and recent findings as well as with further studies showing T_{CM} characteristics to be effectively reactivated.[24–27] Toews *et al* demonstrated that T_{CM}-derived CAR T cells showed an augmented antitumor immunity against neuroblastoma cells under PD-1 blockade and subsequently formed



	Patient characteristics	Tumor characteristics	Blood characteristics
Favors disease control upon PD-1/PD-L1 checkpoint inhibition therapy	No immunosuppression Age ≥70 years ECOG = 0	Only 1 organ involved Brisk infiltrate of effector/memory T cells	Normal LDH Normal CRP
Irrelevant for outcome of PD-1/PD-L1 checkpoint inhibition therapy	Sex Prior chemotherapy Prior radiotherapy	M stage (M1a/b/c) MCPyVirus status PD-L1 status	Neutrophil to Lymphocyte ratio
Favors disease progression upon PD-1/PD-L1 checkpoint inhibition therapy	Immunosuppression Age <70 years ECOG ≥ 1	≥ 2 organs involved No T-cell infiltrate	Elevated LDH Elevated CRP

Figure 5 Schematic overview on relevant clinical and molecular parameters determined before treatment and their predictive value on PD-1/PD-L1 ICI therapy response. CRP, C reactive protein; ECOG, Eastern Cooperative Oncology Group; ICI, immune checkpoint inhibitor; LDH, lactate dehydrogenase.

a resident memory T-cell subset following tumor challenge.[28] In non-small cell lung cancer patients treated with nivolumab, a longer PFS was observed in patients with a high T_{CM}/T_{EFF}-cell ratio in the circulation, suggesting an enrichment of peripheral circulating T_{CM} subpopulations also as a potential positive predictive marker.[29] Similarly, in hepatocellular carcinoma patients, according to midterm clinical trial results, an extended median relapse-free survival was associated with an increased T_{CM}-subpopulation.[22] Moreover, Siddiqui *et al* reported the presence of a $TCF1^{+}PD\text{-}1^{+}CD8^{+}$ T-cell subpopulation in the circulation of melanoma patients and among TILs of primary melanomas. In conclusion, the success of PD-1/PD-L1 ICI therapy seems not to depend on the rejuvenation of differentiated exhausted T cells, but rather on the proliferation of the less-differentiated memory-like $CD8^{+}$ T cells.[30]

Long-lived memory T-cell formation and maintenance are driven by transcription factors like FOXO1, EOMES and TCF1. In particular, TCF1 was identified as the master regulator of genes, inducing serial T-cell reactivation and self-renewal. With respect to the limited predictive value of the presence of granzyme-expressing T cells in the tumor infiltrate, recently a granzyme-positive subpopulation of $CD8^{+}$ T cells associated with age-related dysfunction of the immune system has been described.[31] These cells are characterized by a pronounced tissue-homing capacity and a high clonality, that is, expressing only a limited diversity of TCRs, which might be of particular relevance for MCC, since this tumor affects mainly the elderly population. Indeed, we have shown in previous studies that high clonality of TIL in MCC is both a negative prognostic and predictive biomarker.[5 32]

In conclusion, our results provide a number of clinically well applicable baseline biomarkers associated with PD-1/PD-L1 ICI therapy response in patients suffering from advanced MCC. On a functional level, we confirmed the predominance of T_{CM} among TILs in patients with a favorable ICI therapy response; a factor which can be determined on FFPE tissue.

Author affiliations
[1]Translational Skin Cancer Research, Deutsches Konsortium für Translationale Krebsforschung, Essen, Germany
[2]Department of Dermatology, University Hospital of Essen, Essen, Germany
[3]Department of Dermatology, University Hospital of Lübeck, Lübeck, Germany
[4]Department of Oncology-Pathology, Karolinska University Hospital, Stockholm, Sweden
[5]Department of Dermatology, University Hospital of Tübingen, Tübingen, Germany
[6]Department of Dermatology, University Medical Center Mainz, Mainz, Germany
[7]Department of Dermatology, University Hospital of Heidelberg, Heidelberg, Germany
[8]Deutsches Krebsforschungszentrum, Heidelberg, Germany
[9]Department of Dermatology, Ruhr University Bochum, Bochum, Germany
[10]Department of Dermatology, University Hospital Regensburg, Regensburg, Germany
[11]Department of Dermatology, Elbe Kliniken Buxtehude, Buxtehude, Germany
[12]Department of Dermatology, Saarland University Medical Center, Homburg, Germany
[13]Department of Dermatology-Oncology, University Hospital München, München, Germany
[14]Department of Dermatology, Skin Cancer Center Minden, Minden, Germany
[15]Department of Dermatology, Venereology and Allergology, University Medical Center Mannheim, Mannheim, Germany
[16]Deutsches Konsortium fur Translationale Krebsforschung, Essen, Germany
[17]Department of Pathology, University Hospital Essen, Essen, Germany
[18]Bioinformatics and Computational Biophysics, University of Duisburg-Essen, Duisburg, Germany

Acknowledgements We thank the patients who participated in this study, their families, and the staff members at the clinical centers who cared for them. We like to express our appreciation to Nalini Srinivas for critically proof reading the final version of the manuscript.

Contributors Conceptualisation: IS, SU, LK, H-US, DHa, DHo, and JCB; data curation: IS, SU, and JCB; formal analysis: IS, KH, and DHa; funding acquisition: JCB; investigation: IS, SU, LK, KH, DHa, DHo, and JCB; methodology: IS, SU, LK, DHa, DHo, and JCB; project administration: IS, SU, and JCB; resources: SU, LZ, PT, AM, HBA, LV, UL, TE, CL, JCH, TG, SH, PM, CP, LH, RG, JSU, DS, and JCB; software: IS, KH, and DHa; supervision: SU, and JCB; validation: IS, SU, KH, DHa, DHo, and JCB; visualisation: IS, SU, KH, DHa, DHo, and JCB; writing–original draft: IS, SU, and JCB; writing–review and editing: IS, SU, LK, PT, LV, TE, JCH, TG, RG, KH, DHa, DHo, H-US, DS, and JCB; IS, SU and JCB are guarantors of this study.

Funding This study was funded by the DKTK ED003

Competing interests SU declares research support from Bristol Myers Squibb and Merck Serono; speakers and advisory board honoraria from Bristol Myers Squibb, Merck Sharp & Dohme, Merck Serono, Novartis and Roche, and travel support from Bristol Myers Squibb, and Merck Sharp & Dohme. LZ served as consultant and/or has received honoraria from Roche, Bristol-Myers Squibb, Merck Sharp & Dohme, Novartis, Pierre-Fabre, and Sanofi; Research funding to institution: Novartis; travel support from Merck Sharp & Dohme, Bristol-Myers Squibb, Amgen, Pierre-Fabre, and Novartis, outside the submitted work. PT declares invited speaker's honoraria from Bristol-Myers Squibb, Novartis, MSD, Pierre-Fabre, CureVac, Roche, Kyowa Kirin, Biofrontera, advisory board honoraria from Bristol-Myers Squibb, Novartis, Pierre-Fabre, Merck Serono, Sanofi, Roche, Kyowa Kirin, and travel support from Bristol-Myers Squibb, and Pierre-Fabre. UL declares advisory board honoraria from MSD, Roche, Sanofi, Novartis, Sun Pharma, Almirall Hermal. TE declares consulting fees from BMS, Novartis, Roche, Pierre Fabre, Sanofi; board membership and payment for lectures in speakers bureau from Pierre Fabre, MSD, Roche, BMS, Novartis and Sanofi. CL reports advisory board honoraria from MSD, BMS, Roche, Pierre Fabre, Novartis, Sun Pharma, Sanofi, Kyowa Kirin, Almiral Hermal, Biontech, Merc; speakers fee from MSD, BMS, Roche, Pierre Fabre, Novartis, Sun Pharma, Sanofi, Merck; travel reimbusment from MSD, BMS, Roche, Pierre Fabre, Novartis, Sun Pharma, Sanofi, Kyowa Kirin, Almiral Hermal, Biontech and Merck. JCH declares research support from Bristol Myers Squibb; advisory board honoraria from Pierre Fabre, Sanofi, Sun Pharma and Merck Sharp & Dome; speakers honoraria from Bristol Myers Squibb, Merck Sharp & Dohme, Novartis, Roche, Sanofi and Almirall and travel support from Pierre Fabre. TG reports receiving speakers and/or advisory board honoraria from BMS, Sanofi-Grenzyme, MSD, Novartis Pharma, Roche, Abbvie, Almirall, Janssen Lilly, Pfizer, Pierre Fabre; speakers and/or advisory board honoraria from BMS, Sanofi-Grenzyme, MSD, Novartis, Pharma, Roche, Abbvie, Almirall, Janssen Lillly, Pfizer, Pierre Fabre. SH declares advisory boards honoraria from Pierre Fabre, MSD, BMS, Novartis, Sanofi. PM reports board membership and payment for lectures in speakers bureau from Pierre Fabre, GSK, MSD, Merck Germany, Roche, BMS, Novartis, Sanofi. P. Mohr reports research support (to institution): Bristol-MyersSquibb, Novartis, MSD. Honoraria for lectures (personally): Pierre Fabre, GSK, MSD, Merck Germany, Roche, BMS, Novartis, Sanofi, Amgen, SUN-Pharma, Roche Pharma, Bristol-MyersSquibb, Novartis, MSD, Almirall-Hermal, Amgen, Merck-Serono, Bayer, Pierre-Fabre, Sanofi. Honoraria for advisory boards: Pierre Fabre, GSK, MSD, Merck Germany, Roche, BMS, Novartis, Sanofi, Beiersdorf, Almiral-Hermal, AmgenBayersdorf, Roche Pharma, Bristol-MyersSquibb, Novartis, MSD, Almirall-Hermal, Amgen, Pierre-Fabre, Merck-Serono, SUN-Pharma, SUN, Merck-Serono, Sanofi. CP received honoraria (speaker honoraria or honoraria as a consultant) and travel support from Novartis, BMS, Roche, Merck Serono, MSD, Celgene, AbbVie, AMGEN, SUNPHARMA, Allergy Therapeutics and LEO. LMH served as consultant and/or has received honoraria from Amgen, BMS, Curevac, MSD, Novartis, Pierre-Fabre, Roche, Sanofi and Sun Pharma, outside the submitted work. Research funding to institution: Novartis. R. Gutzmer reports research support from Pfizer, Johnson & Johnson, Novartis, Amgen, MerckSerono, SUN Pharma; honoraria for lectures from Roche Pharma, Bristol-MyersSquibb, Novartis, MSD, Almirall-Hermal, Amgen, Merck-Serono, SUN, Pierre-Fabre, Sanofi, SUN Pharma, Bayer; honoraria for advice from Roche Pharma, Bristol-MyersSquibb, Novartis, MSD, Almirall-Hermal, Amgen, Pierre-Fabre, Merck-Serono, 4SC, Incyte, SUN Pharma, Sanofi, Pfizer. JSU is on the advisory board or has received honoraria and travel support from Amgen, Bristol Myers Squibb, GSK, LeoPharma, Merck Sharp and Dohme, Novartis, Pierre Fabre, Roche, Sanofi outside the submitted work. H-US is an employee of Targos Molecular Pathology Inc. and reports research support from Novartis Oncology and received honoraria from MSD, BMS, Roche Pharma, Novartis Oncology, AstraZeneca, Eisai, Takeda, Molecular Health, outside of the submitted work. DS reports personal fees from Amgen, GSK, BMS, Novartis, Roche, Merck, Astra Zeneca, Merck-Serono, Pfizer, Incyte, Array Pierre Fabre, Sanofi Genzyme, Regeneron, 4Sc, InFlaRx, Neracare, Ultimovacs, SunPharma, Philogen, Immunocore, Sandoz-Hexal outside the submitted work. JCB is receiving speaker's bureau honoraria from Amgen, Pfizer, MerckSerono, Recordati and Sanofi; is a paid consultant/advisory board member/DSMB member for Almirall, Boehringer Ingelheim, InProTher, ICON, MerckSerono, Pfizer, 4SC, and Sanofi/Regeneron. His group receives research grants from Bristol-Myers Squibb, Merck Serono, HTG, IQVIA, and Alcedis.

Patient consent for publication Not applicable.

Ethics approval This study was approved by Ethics committee of the University Duisburg-Essen (11-4715; 17-7538-BO) Ethics board of Stockholm (Dnr 2010/1092-31/3).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request. All data relevant to the study are included in the article or uploaded as online supplemental information. All the raw data used within this manuscript are available on reasonable request to the corresponding author.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iDs

Ivelina Spassova http://orcid.org/0000-0003-4663-7966
Selma Ugurel http://orcid.org/0000-0002-9384-6704
Patrick Terheyden http://orcid.org/0000-0002-5894-1677
Jürgen Christian Becker http://orcid.org/0000-0001-9183-653X

REFERENCES

1 Becker JC, Stang A, DeCaprio JA, *et al*. Merkel cell carcinoma. *Nat Rev Dis Primers* 2017;3:17077.
2 Nghiem PT, Bhatia S, Lipson EJ, *et al*. PD-1 blockade with pembrolizumab in advanced Merkel-cell carcinoma. *N Engl J Med* 2016;374:2542–52.
3 Topalian SL, Bhatia S, Amin A, *et al*. Neoadjuvant nivolumab for patients with resectable Merkel cell carcinoma in the CheckMate 358 trial. *J Clin Oncol* 2020;38:2476–87.
4 Kacew AJ, Dharaneeswaran H, Starrett GJ, *et al*. Predictors of immunotherapy benefit in Merkel cell carcinoma. *Oncotarget* 2020;11:4401–10.
5 Spassova I, Ugurel S, Terheyden P, *et al*. Predominance of central memory T cells with high T-cell receptor repertoire diversity is associated with response to PD-1/PD-L1 inhibition in Merkel cell carcinoma. *Clin Cancer Res* 2020;26:2257–67.
6 D'Angelo SP, Russell J, Lebbe C. Efficacy and safety of first-line avelumab treatment in patients with stage IV metastatic Merkel cell carcinoma: a preplanned interim analysis of a clinical trial. *JAMA Oncol* 2018;4:1–5.
7 Eisenhauer EA, Therasse P, Bogaerts J, *et al*. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer* 2009;45:228–47.
8 Shen R, Postow MA, Adamow M, *et al*. LAG-3 expression on peripheral blood cells identifies patients with poorer outcomes after immune checkpoint blockade. *Sci Transl Med* 2021;13:eabf5107.
9 Bürkner P-C, Vuorre M. Ordinal regression models in psychology: a tutorial. *Adv Methods Pract Psychol Sci* 2019;2:77–101.
10 Thomas DC. Introduction: Bayesian models and Markov chain Monte Carlo methods. *Genet Epidemiol* 2001;21:S660–1.
11 Bürkner PC. Advanced Bayesian multilevel modeling with the R package brms. *R Journal* 2018;180:1–15.
12 Bürkner PC. brms: an R package for Bayesian multilevel models using Stan. *J Stat Softw* 2017;80:1–28.
13 Buuren Svan, Groothuis-Oudshoorn K. mice : Multivariate Imputation by Chained Equations in *R*. *J Stat Softw* 2011;45:1–67.
14 Collins GS, Reitsma JB, Altman DG, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015;162:55–63.
15 Henson SM, Riddell NE, Akbar AN. Properties of end-stage human T cells defined by CD45RA re-expression. *Curr Opin Immunol* 2012;24:476–81.
16 Walker JW, Lebbé C, Grignani G, *et al*. Efficacy and safety of avelumab treatment in patients with metastatic Merkel cell carcinoma: experience from a global expanded access program. *J Immunother Cancer* 2020;8:e000313.
17 Kugel CH, Douglass SM, Webster MR, *et al*. Age correlates with response to anti-PD1, reflecting age-related differences in intratumoral effector and regulatory T-cell populations. *Clin Cancer Res* 2018;24:5347–56.
18 Johnstone J, Parsons R, Botelho F, *et al*. T-Cell phenotypes predictive of frailty and mortality in elderly nursing home residents. *J Am Geriatr Soc* 2017;65:153–9.
19 De Maeyer RPH, Chambers ES. The impact of ageing on monocytes and macrophages. *Immunol Lett* 2021;230:1–10.
20 Knepper TC, Montesion M, Russell JS, *et al*. The genomic landscape of Merkel cell carcinoma and Clinicogenomic biomarkers of response to immune checkpoint inhibitor therapy. *Clin Cancer Res* 2019;25:5961–71.
21 Lanzavecchia A, Sallusto F. Regulation of T cell immunity by dendritic cells. *Cell* 2001;106:263–6.
22 Liu Q, Sun Z, Chen L. Memory T cells: strategies for optimizing tumor immunotherapy. *Protein Cell* 2020;11:549–64.
23 Sallusto F, Geginat J, Lanzavecchia A. Central memory and effector memory T cell subsets: function, generation, and maintenance. *Annu Rev Immunol* 2004;22:745–63.
24 Im SJ, Hashimoto M, Gerner MY, *et al*. Defining CD8+ T cells that provide the proliferative burst after PD-1 therapy. *Nature* 2016;537:417–21.
25 Kurtulus S, Madi A, Escobar G, *et al*. Checkpoint blockade immunotherapy induces dynamic changes in PD-1−CD8+ tumor-infiltrating T cells. *Immunity* 2019;50:181–94.
26 Miron M, Kumar BV, Meng W, *et al*. Human Lymph Nodes Maintain TCF-1 ^{hi} Memory T Cells with High Functional Potential and Clonal Diversity throughout Life. *J Immunol* 2018;201:2132–40.
27 Sade-Feldman M, Yizhak K, Bjorgaard SL, *et al*. Defining T cell states associated with response to checkpoint immunotherapy in melanoma. *Cell* 2018;175:998–1013.
28 Toews K, Grunewald L, Schwiebert S, *et al*. Central memory phenotype drives success of checkpoint inhibition in combination with CAR T cells. *Mol Carcinog* 2020;59:724–35.
29 Manjarrez-Orduño N, Menard LC, Kansal S, *et al*. Circulating T cell subpopulations correlate with immune responses at the tumor site and clinical response to PD1 inhibition in non-small cell lung cancer. *Front Immunol* 2018;9:1613.
30 Siddiqui I, Schaeuble K, Chennupati V, *et al*. Intratumoral Tcf1^{+}PD-1^{+}CD8^{+} T Cells with Stem-like Properties Promote Tumor Control in Response to Vaccination and Checkpoint Blockade Immunotherapy. *Immunity* 2019;50:195–211.
31 Mogilenko DA, Shpynov O, Andhey PS, *et al*. Comprehensive Profiling of an Aging Immune System Reveals Clonal GZMK^{+} CD8^{+} T Cells as Conserved Hallmark of Inflammaging. *Immunity* 2021;54:99–115.
32 Farah M, Reuben A, Spassova I, *et al*. T-Cell repertoire in combination with T-cell density predicts clinical outcomes in patients with Merkel cell carcinoma. *J Invest Dermatol* 2020;140:2146–56.

Supplementary Table S1. Detailed patient characteristics at baseline, during anti-PD-1/PD-L1 therapy, and follow-up.

Pat ID	Age (years) at baseline	Sex	ECOG at baseline	Immuno-suppression	Serum LDH at baselin e	Serum CRP at baseline	neutro lympho ratio at baseline	Locali-sation of primary	M stage at therapy start	Organs involved	MCPyV status (tumor)	PD-L1 expression (tumor)	Previous chemo-therapy	Previous radio-therapy	Immune Checkpoint Inhibitors	Best Response	Trial of partici-pation	Year of treatment
1	86	M	0	No	normal	elevated	≥4	head + neck	M0	1	nd	nd	No	No	Avelumab	PD	No	2018
2	81	M	1	Yes (CLL; Prostate Cancer)	elevated	elevated	nd	extremities	M1a	2	positive	negative	No	No	Avelumab	PD	No	2018
3	79	F	1	No	normal	normal	<4	head + neck	M1c	1	positive	negative	Yes	Yes	Pembrolizumab	PD	No	2016
4	70	M	1	No	elevated	elevated	<4	extremities	M0	2	positive	nd	No	No	Avelumab	PD	No	2019
5	83	M	1	No	normal	normal	<4	head + neck	M1a	2	negative	positive	No	No	Avelumab	PD	No	2018
6	68	F	1	Yes (azathioprine, corticosteroids (rheumatoid arthritis, systemic lupus erythematodes, Hashimoto thyroiditis))	elevated	normal	≥4	extremities	M1c	2	positive	negative	No	Yes	Avelumab	PD	No	2017
7	66	M	0	No	normal	elevated	<4	trunk	M1a	1	nd	nd	No	No	Avelumab	PD	Yes	2017
8	62	M	0	No	elevated	normal	<4	unknown	M0	1	positive	positive	No	No	Avelumab	PD	Yes	2017
9	76	M	1	No	elevated	elevated	≥4	unknown	M1a	1	nd	nd	Yes	No	Avelumab	PD	Yes	2015
10	71	F	4	Yes (tacrolimus, corticosteroids (kidney transplantation))	elevated	elevated	<4	head + neck	M1c	2	negative	positive	No	No	Nivolumab	PD		2015
11	76	F	0	No	normal	nd	≥4	trunk	M1c	3	nd	nd	No	Yes	Pembrolizumab	PD	No	2017
12	75	M	1	Yes (CLL)	normal	nd	nd	unknown	M1b	2	positive	nd	No	No	Pembrolizumab	PD	No	2018
13	77	F	1	No	elevated	nd	nd	unknown	M1b	2	nd	nd	No	No	Pembrolizumab	PD	Yes	2019
14	86	M	0	No	elevated	nd	nd	unknown	M1c	3	nd	nd	No	No	Pembrolizumab	PD	Yes	2019
15	67	F	1	No	elevated	elevated	≥4	head + neck	M1a	1	positive	positive	No	Yes	Pembrolizumab	PD	No	2016
16	66	F	1	No	elevated	elevated	<4	unknown	M0	2	nd	nd	Yes	Yes	Pembrolizumab	PD	No	2016
17	72	M	1	No	elevated	elevated	<4	extremities	M1a	1	positive	negative	Yes	Yes	Pembrolizumab	PD	No	2016
18	62	M	0	No	elevated	elevated	≥4	extremities	M1c	2	nd	nd	No	Yes	Avelumab	PD	Yes	2016
19	79	M	0	No	elevated	elevated	<4	trunk	M1c	3	nd	nd	No	No	Avelumab	PD	No	2018
20	68	M	1	Yes (azathioprine (myasthenia gravis)	elevated	elevated	nd	head + neck	M0	1	positive	negative	No	Yes	Nivolumab	PD	No	2017
21	64	M	0	Yes (NH-Lymphom)	normal	nd	<4	extremities	M1c	4	nd	nd	No	Yes	Avelumab	PD	No	2018
22	77	M	0	Yes (CLL)	elevated	normal	<4	extremities	M1a	1	nd	nd	No	No	Avelumab	PD	No	2019

23	83	M	1	No	elevated	elevated	≥4	extremities	M0	1	nd	nd	Yes	No	Avelumab	PD	Yes	2015
24	75	M	2	No	normal	elevated	≥4	extremities	M1c	2	negative	nd	No	Yes	Avelumab	PD	No	2018
25	56	M	1	No	normal	nd	<4	extremities	M1a	2	nd	nd	No	Yes	Pembrolizumab	PD	No	2016
26	74	M	0	No	elevated	elevated	≥4	extremities	M1c	5	nd	nd	Yes	Yes	Avelumab	PD	Yes	2015
27	75	F	1	No	elevated	normal	≥4	unknown	M1a	1	positive	negative	Yes	No	Pembrolizumab	PD	No	2016
28	82	F	1	No	elevated	normal	<4	extremities	M1a	2	nd	positive	No	Yes	Avelumab	PD	No	2018
29	83	M	3	No	elevated	elevated	≥4	unknown	M1c	2	nd	nd	No	Yes	Avelumab	PD	No	2019
30	59	M	0	Yes (CLL)	normal	normal	<4	extremities	M1a	2	positive	positive	No	No	Nivolumab	PD	No	2017
31	66	F	2	No	elevated	elevated	<4	trunk	M1a	1	nd	nd	No	No	Avelumab	PD	Yes	2017
32	56	F	0	No	elevated	normal	<4	extremities	M1c	3	nd	nd	No	Yes	Avelumab	PD	Yes	2016
33	78	F	0	No	nd	nd	nd	trunk	M1c	3	nd	nd	Yes	Yes	Avelumab	PD	No	2017
34	37	M	0	No	normal	elevated	<4	trunk	M1b	2	nd	positive	Yes	Yes	Avelumab	PD	Yes	2015
35	85	M	1	No	elevated	elevated	<4	extremities	M1c	4	positive	positive	No	No	Pembrolizumab	PD	No	2017
36	66	M	1	Yes (CLL)	elevated	elevated	nd	trunk	M1a	1	negative	positive	No	No	Pembrolizumab	PD	No	2018
37	53	M	1	Yes (MTX in clippers syndrome and steroids)	elevated	elevated	≥4	extremities	M1c	3	nd	nd	No	Yes	Avelumab	PD	No	2018
38	80	M	0	No	elevated	elevated	≥4	unknown	M1c	2	nd	nd	No	Yes	Avelumab	PD	Yes	2017
39	80	M	0	Yes (rheumatoid arthritis (until 11/2016 MTX + Ankinra, prednisolone 5mg daily))	elevated	elevated	≥4	head + neck	M1c	5	nd	nd	Yes	Yes	Pembrolizumab	PD		2016
40	69	M	0	Yes (CML, osteomyelofibrosis)	elevated	elevated	≥4	extremities	M1c	2	nd	nd	No	No	Avelumab	PD		2017
41	68	M	0	Yes (CLL)	elevated	elevated	nd	unknown	M1a	1	nd	nd	Yes	No	Pembrolizumab	SD	No	2017
42	77	M	1	Yes (CLL, melanoma)	normal	elevated	≥4	head + neck	M1c	1	negative	positive	Yes	Yes	Pembrolizumab	SD	No	2016
43	75	M	0	No	elevated	elevated	<4	extremities	M0	1	nd	nd	Yes	No	Avelumab	SD	No	2018
44	78	M	1	No	elevated	elevated	<4	extremities	M1c	3	nd	negative	Yes	No	Avelumab	SD	Yes	2015
45	48	F	0	No	normal	normal	<4	unknown	M1a	1	positive	positive	Yes	Yes	Pembrolizumab	SD	No	2017
46	76	M	0	No	normal	normal	<4	head + neck	M1c	3	negative	positive	No	No	Avelumab	SD	No	2018
47	72	F	0	No	elevated	elevated	≥4	head + neck	M1a	1	positive	negative	Yes	Yes	Nivolumab	SD	Yes	2015
48	60	F	0	No	normal	nd	<4	extremities	M1c	4	nd	nd	No	No	Avelumab	SD	No	2019
49	83	M	1	No	elevated	elevated	<4	unknown	M1c	4	nd	nd	Yes	Yes	Pembrolizumab	SD	No	2014
50	82	M	0	No	normal	elevated	<4	head + neck	M0	1	nd	nd	Yes	No	Avelumab	SD	Yes	2016

51	63	M	0	No	elevated	nd	<4	extremities	M1c	4	nd	nd	No	Yes	Avelumab	SD	No	2018
52	90	F	2	No	elevated	elevated	nd	head + neck	M0	1	nd	nd	No	No	Avelumab	SD	No	2017
53	80	F	0	Yes (non-Hodgkin lymphoma (cyclophosphamide, doxorubicine, vincristine, prednisone); anal carcinoma)	normal	normal	<4	head + neck	M1c	2	negative	negative	No	No	Nivolumab	SD	No	2017
54	78	M	0	Yes (CLL, melanoma)	elevated	normal	<4	extremities	M1c	2	positive	nd	No	Yes	Pembrolizumab	SD	No	2017
55	79	M	0	No	elevated	elevated	≥4	extremities	M1c	3	nd	nd	No	Yes	Pembrolizumab	SD		2017
56	83	F	1	Yes (multiple myeloma, polycythemia vera)	elevated	elevated	<4	extremities	M1c	2	nd	nd	No	Yes	Avelumab	SD		2017
57	73	M	1	No	elevated	elevated	≥4	trunk	M1c	4	negative	positive	Yes	Yes	Avelumab	SD	Yes	2016
58	68	M	0	No	normal	nd	<4	unknown	M1a	1	nd	nd	No	No	Avelumab	SD	No	2018
59	70	M	0	No	elevated	elevated	<4	extremities	M1a	1	positive	positive	Yes	Yes	Pembrolizumab	SD	No	2016
60	69	F	2	No	normal	normal	<4	head + neck	M1c	2	nd	nd	No	No	Avelumab	SD	No	2018
61	82	M	1	No	normal	nd	<4	head + neck	M1c	3	nd	nd	No	No	Avelumab	PR	No	2017
62	86	M	nd	No	normal	normal	<4	extremities	M0	1	nd	nd	No	No	Avelumab	PR	No	2018
63	71	M	0	No	elevated	normal	<4	extremities	M1a	2	positive	positive	No	Yes	Pembrolizumab	PR	No	2017
64	89	M	0	Yes (non-small cell lung cancer; prostate cancer)	normal	normal	<4	trunk	M1c	3	positive	negative	No	Yes	Pembrolizumab	PR	No	2017
65	65	M	0	No	normal	nd	<4	extremities	M1a	1	positive	negative	No	Yes	Nivolumab	PR	Yes	2016
66	73	M	0	No	normal	normal	≥4	extremities	M1a	1	nd	nd	No	Yes	Nivolumab	PR	No	2015
67	90	M	1	No	elevated	elevated	≥4	head + neck	M1c	2	nd	nd	No	Yes	Pembrolizumab	PR		2017
68	74	M	1	No	elevated	elevated	≥4	extremities	M1c	2	positive	negative	Yes	Yes	Pembrolizumab	PR	No	2015
69	57	F	0	No	normal	normal	<4	extremities	M1a	1	nd	nd	No	No	Avelumab	PR	No	2018
70	68	M	0	No	normal	normal	≥4	head + neck	M1c	3	negative	negative	Yes	No	Avelumab	PR	Yes	2015
71	59	M	0	no	normal	nd	nd	unknown	M1b	3	nd	nd	Yes	Yes	Pembrolizumab	PR	No	2017
72	55	M	0	No	elevated	elevated	≥4	unknown	M1c	2	nd	nd	No	Yes	Pembrolizumab	PR	No	2016
73	85	M	0	No	elevated	nd	<4	trunk	M1c	1	nd	nd	No	Yes	Pembrolizumab	PR	No	2016
74	76	M	0	No	elevated	elevated	<4	extremities	M1c	2	nd	positive	Yes	No	Pembrolizumab	PR	No	2015
75	86	M	1	No	elevated	elevated	≥4	trunk	M1c	2	positive	positive	No	No	Pembrolizumab	PR		2017

76	84	M	1	No	elevated	normal	≥4	trunk	M1c	4	positive	positive	No	Yes	Avelumab	PR	No	2018
77	61	M	0	No	elevated	normal	<4	head + neck	M1c	3	nd	nd	No	Yes	Avelumab	PR	Yes	2016
78	76	M	0	No	normal	elevated	<4	unknown	M1a	1	nd	nd	No	Yes	Pembrolizumab	PR		2017
79	48	M	1	Yes (promyelocytic leukemia)	elevated	elevated	<4	trunk	M0	1	positive	negative	No	Yes	Avelumab	PR	No	2018
80	85	M	1	No	elevated	nd	nd	unknown	M1c	3	positive	nd	No	Yes	Pembrolizumab	PR	Yes	2019
81	50	M	1	No	elevated	elevated	≥4	extremities	M0	1	positive	positive	Yes	No	Nivolumab	PR	No	2016
82	71	F	1	No	normal	nd	nd	unknown	M1a	1	nd	nd	No	No	Pembrolizumab	PR	Yes	2019
83	52	M	0	No	elevated	elevated	≥4	trunk	M1a	1	nd	nd	No	Yes	Avelumab	PR	No	2018
84	82	M	1	Yes (renal cell carcinoma, urothelial carcinoma)	elevated	elevated	<4	extremities	M1a	1	nd	negative	No	Yes	Avelumab	PR		2018
85	76	M	1	Yes (CLL)	nd	nd	nd	unknown	M1a	1	nd	nd	Yes	No	Nivolumab	PR	No	2016
86	82	F	2	No	elevated	nd	nd	unknown	M1c	3	nd	nd	No	No	Pembrolizumab	PR	No	2017
87	82	M	0	No	elevated	nd	nd	unknown	M1c	2	positive	nd	No	No	Pembrolizumab	PR	No	2019
88	77	M	0	No	normal	elevated	≥4	head + neck	M1c	1	nd	nd	No	Yes	Pembrolizumab	PR		2016
89	75	F	1	No	elevated	elevated	≥4	extremities	M1b	1	nd	nd	No	No	Avelumab	PR		2018
90	87	M	1	No	elevated	elevated	<4	extremities	M0	1	nd	nd	No	No	Avelumab	PR	No	2017
91	65	M	0	No	normal	normal	<4	head + neck	M1c	1	nd	nd	No	No	Avelumab	CR	Yes	2017
92	59	M	1	No	normal	normal	<4	unknown	M1c	2	nd	positive	Yes	Yes	Pembrolizumab	CR	No	2016
93	96	M	1	No	normal	nd	≥4	trunk	M0	1	nd	nd	No	No	Avelumab	CR	No	2018
94	64	M	0	No	normal	elevated	<4	trunk	M1c	1	nd	nd	No	Yes	Pembrolizumab	CR	No	2016
95	61	F	0	Yes (azathioprine (myasthenia gravis))	elevated	elevated	<4	trunk	M1c	1	nd	nd	No	Yes	Pembrolizumab	CR	No	2016
96	69	M	1	No	elevated	normal	<4	extremities	M1a	2	positive	negative	Yes	Yes	Avelumab	CR	Yes	2016
97	79	M	0	No	elevated	normal	≥4	extremities	M1a	1	nd	nd	No	Yes	Nivolumab	CR	Yes	2016
98	57	F	0	No	normal	nd	≥4	unknown	M1c	3	positive	negative	No	Yes	Nivolumab	CR	Yes	2015
99	76	M	0	No	elevated	elevated	<4	extremities	M1a	1	negative	nd	No	No	Pembrolizumab	CR	No	2017
100	72	F	0	No	normal	elevated	≥4	extremities	M1a	1	nd	positive	Yes	Yes	Avelumab	CR	No	2017
101	71	M	0	No	normal	normal	<4	extremities	M0	2	nd	nd	No	Yes	Avelumab	CR	No	2018
102	79	F	0	No	elevated	normal	<4	trunk	M1a	2	positive	nd	No	Yes	Avelumab	CR	Yes	2016
103	83	F	1	No	elevated	elevated	≥4	unknown	M1a	2	nd	nd	No	No	Pembrolizumab	CR		2017
104	80	M	0	No	normal	normal	<4	head + neck	M1c	1	nd	nd	No	Yes	Avelumab	CR		2017

105	60	M	0	No	elevated	elevated	nd	extremities	M1c	5	nd	nd	Yes	Yes	Pembrolizumab	CR	No	2016
106	82	F	0	No	nd	nd	nd	extremities	M1c	2	nd	nd	No	No	Avelumab	CR	No	2018
107	62	M	0	No	normal	nd	nd	extremities	M1c	1	nd	nd	No	No	Pembrolizumab	CR	No	2016
108	75	F	1	No	normal	nd	nd	head + neck	M1c	2	nd	nd	Yes	No	Pembrolizumab	CR	No	2017
109	83	F	0	No	normal	nd	nd	head + neck	M0	1	nd	nd	No	No	Avelumab	CR	No	2018
110	83	M	0	No	normal	normal	nd	head + neck	M0	1	nd	nd	No	No	Avelumab	CR	No	2018
111	79	M	0	No	nd	nd	nd	trunk	M1a	1	nd	nd	No	Yes	Avelumab	CR	No	2017
112	77	M	0	No	elevated	nd	nd	unknown	M1c	2	positive	nd	No	No	Nivolumab	CR	No	2016
113	77	F	0	No	elevated	nd	nd	unknown	M1a	1	positive	nd	No	No	Pembrolizumab	CR	No	2017
114	64	M	0	No	normal	nd	nd	unknown	M1a	1	positive	nd	No	No	Nivolumab	CR	No	2017

Characteristics of total patient cohort (n=118). Best response is related to the anti-PD-1/PD-L1 therapy. Patients are sorted according to their change in sum of longest diameters of target lesions from baseline to best response (see **Figure 1**). Abbreviations: M – male; F – female; CLL – chronic lymphocytic leukemia; CML – chronic myelogenous leukemia; CR – complete response, PR – partial response; SD – stable disease; PD – progressive disease; nd – no data available.

Supplementary Table S2. Antibodies and procedures used for multiplex immunohistochemical staining

Position	Antibody	Clone/Company	Dilution	Incubation	AG[1] retrieval	TSA[3] dye
Panel 1						
1	CD4	PerkinElmer	1:50	30 min	AR[2] 9	520
2	CD8	PerkinElmer	1:100	30 min	AR 9	570
3	CD20	PerkinElmer	1:200	30 min	AR 6	540
4	FoxP3	PerkinElmer	1:400	30 min	AR 6	620
5	CD68	PerkinElmer	1:1000	30 min	AR 6	650
6	SYN	SP11/Abcam	1:1000	Over night	AR 6	690
Panel 2						
1	CD27	EPR8569/Abcam	1:2000	30 min	AR 9	520
2	GZMB	ab4059/Abcam	1:100	30 min	AR 6	570
3	TCF7	C63D9/Cell Signaling	1:100	30 min	ÂR 6	540
4	CD45RA	4KB5/Santa Cruz	1:500	30 min	AR 6	620
5	CD45R0	UCHL1/Novus Bio.	1:1000	30 min	AR 9	650
6	SYN	SP11/Abcam	1:1000	Over night	AR 6	690

[1]AG: antigen; [2]AR: antigen retrieval buffer with pH 6 (AR6) and pH 9 (AR9); SYN: synaptophysin; [3]TSA: Tyramide Signal Amplification

Supplementary Table S3. Markers for detection/quantification of immune cell types in MCC tissue.

Cell type	**Used markers for detection/quantification**
Leukocytes	CD45RA(+) or CD45RO(+)
Regulatory T cells	CD4(+)FoxP3(+)
Central memory T cells	CD27(+)TCF1(+)CD45RO(+)
Effector T cells	GZMB(+)CD45RA(+)
Monocytes/macrophages	CD68(+)
B cells	CD20(+)

Supplementary Table S4. Mean percentage of positively stained cells to total cell number per analysed area for each analyzed MCC samples. For quantification analysis three randomly chosen tissue regions at the juxta-tumoral area as well as in the intra-tumoral area were processed in a semi-automatic fashion by InForm Tissue Analysis software.

Response to ICI	total cell number (in three tissue regions)		% to total cell number for all three fields											
			Ø CD4(+)		ØFoxP3(+)		Ø CD4(+)FoxP3(+)		Ø CD8(+)		Ø CD20(+)		Ø CD68(+)	
	Juxta-tumoral	Intra-tumoral	Juxta-tumoral	Intra-tumoral	Juxta-tumoral	Intra-tumoral	Juxta-tumoral	Intra-tumoral	Juxta-tumoral	Intra-tumoral	Juxta-tumoral	Intra-tumoral	Juxta-tumoral	Intra-tumoral
PR	7034	6042	1,94	0,83	1,98	0,69	1,97	0,68	7,97	4,63	5,40	2,60	3,20	1,99
CR	7019	6710	0,94	14,22	6,26	3,91	3,27	2,13	30,20	20,57	16,77	13,01	2,05	1,54
SD	5845	6390	0,63	0,90	1,76	1,32	0,03	0,03	10,50	12,79	4,41	4,14	2,94	3,91
PR	5443	5167	1,52	1,38	1,60	1,23	0,49	0,55	12,46	10,52	2,76	2,11	4,14	4,47
PR	5160	7304	3,60	0,09	2,75	0,26	1,47	0,19	3,58	0,23	2,01	0,03	3,51	0,64
SD	6414	8940	0,16	0,02	1,76	1,32	0,43	0,00	2,58	0,20	1,27	0,18	1,89	0,29
CR	3196	4847	1,02	1,05	2,61	3,38	0,83	0,83	3,30	7,27	0,21	0,40	1,55	3,60
PR	4703	7812	5,37	0,00	0,00	0,00	0,00	0,00	31,20	7,33	0,00	0,00	3,73	0,30
PR	3604	4782	0,50	2,25	0,00	0,00	0,00	0,00	11,92	9,04	0,40	1,00	1,75	1,00
CR	4434	5039	0,25	0,30	0,00	0,00	0,00	0,00	14,21	2,65	0,00	0,00	3,50	0,45
PR	3005	2707	9,30	3,90	0,90	1,10	0,90	1,10	42,62	15,88	4,65	0,22	4,98	3,56
PD	-	5114	-	0,00	-	0,15	-	0,00	-	0,96	-	0,00	-	0,60
PD	4366	5065	0,11	0,00	0,13	0,00	0,08	0,00	0,00	0,07	0,00	0,00	0,28	0,39
PD	4302	4052	1,46	0,33	2,53	1,79	0,81	0,23	3,30	0,63	0,20	0,08	2,27	1,57
PD	3184	4720	2,86	2,66	3,35	2,30	2,80	2,02	6,63	3,12	6,04	2,06	1,76	1,12
PD	7849	6980	1,32	1,91	0,49	0,24	0,27	0,14	6,23	7,90	7,47	4,32	1,18	2,10
PD	2644	6392	11,20	3,45	0,00	0,00	0,00	0,00	8,00	7,88	0,10	1,05	18,30	9,10
PD	3518	3648	2,00	0,10	0,00	0,00	0,00	0,00	9,04	6,15	0,00	0,00	1,73	0,35
PD	4531	5983	0,15	0,05	0,00	0,00	0,00	0,00	19,50	6,50	6,25	0,00	7,95	0,40
PD	4227	5709	0,40	0,00	0,00	0,00	0,00	0,00	2,59	0,96	0,00	0,00	5,50	0,20
PD	3218	3146	0,20	0,55	0,00	0,00	0,00	0,00	8,00	1,05	0,10	0,10	4,45	0,75

Supplementary Table S5. Mean percentage of positively stained cells to total cell number per analysed area for each analyzed MCC samples. For quantification analysis of three randomly chosen tissue regions at the juxta-tumoral area as well as in the intra-tumoral area were processed in a semi-automatic fashion by InForm Tissue Analysis software.

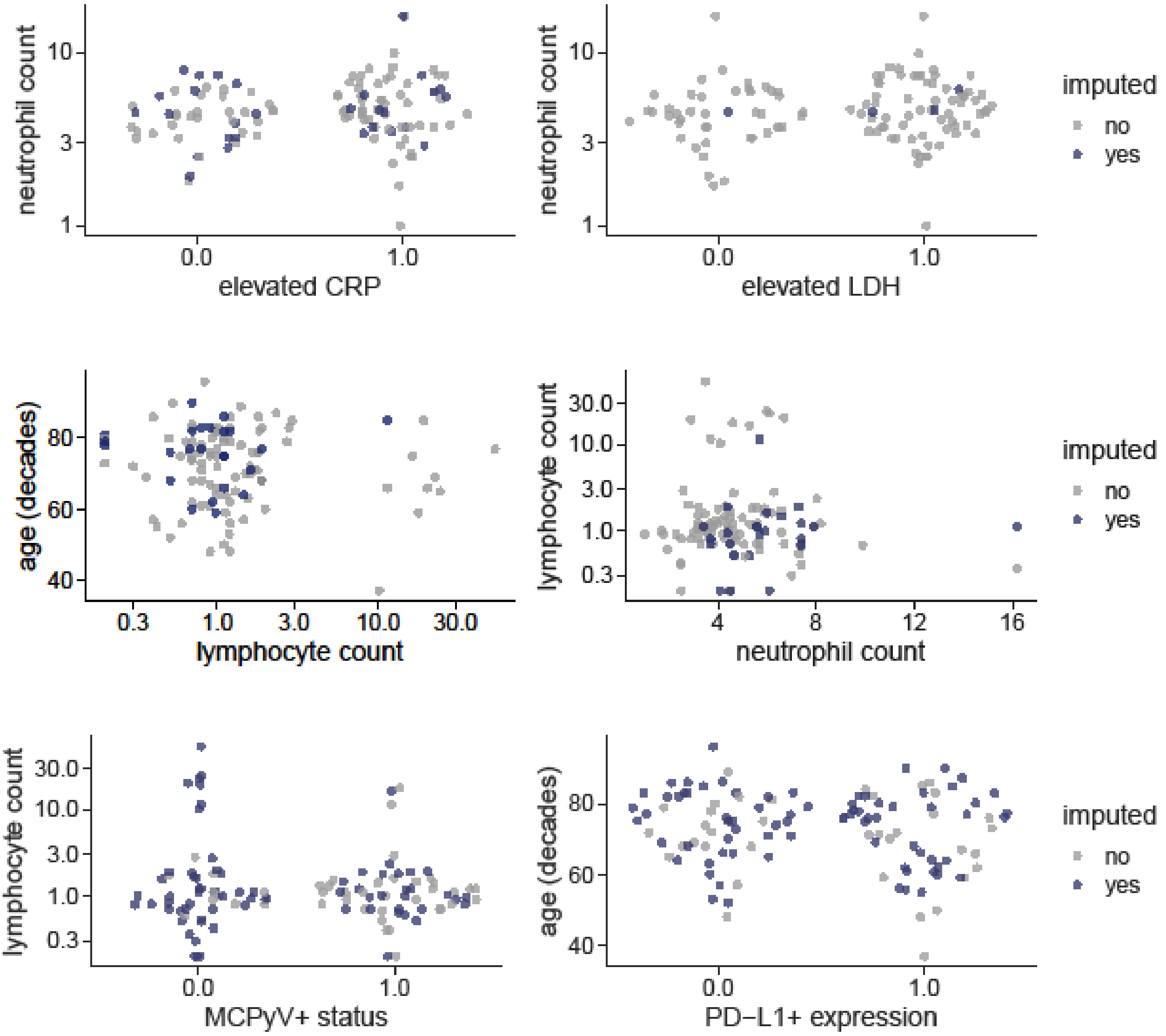
Response to ICI	total cell number (in three tissue regions)		% to total cell number for all three fields													
			Ø CD45RO(+)		Ø CD45RA(+)		Ø CD27(+)		Ø TCF7(+)		Ø CD27(+)TCF7(+) CD45RO(+)		Ø GZMB(+)		Ø GZMB(+)CD45RA(+)	
	Juxta-tumoral	Intra-tumoral	Juxta-tumoral	Intra-tumoral	Juxta-tumoral	Intra-tumoral	Juxta-tumoral	Intra-tumoral	Juxta-tumoral	Intra-tumoral	Juxta-tumoral	Intra-tumoral	Juxta-tumoral	Intra-tumoral	Juxta-tumoral	Intra-tumoral
PR	7182	6571	26,33	9,19	3,20	1,94	13,63	11,41	7,30	2,62	1,39	0,24	0,17	0,07	0,00	0,00
CR	6943	7694	15,92	10,72	5,83	2,14	5,54	4,68	3,97	2,33	0,65	0,35	4,19	3,77	0,99	0,64
SD	6992	6464	0,06	34,64	25,87	19,80	11,97	41,78	5,57	23,73	0,00	11,98	0,30	3,40	0,10	1,73
PR	5727	7196	46,29	68,40	48,63	62,30	19,53	17,59	66,97	64,60	17,09	15,29	6,40	4,39	3,50	2,63
PR	5665	6431	30,33	29,15	3,25	4,46	48,10	17,83	8,07	8,45	2,25	2,69	2,58	1,53	0,11	0,08
SD	4908	6575	22,27	0,87	7,30	0,60	12,76	0,60	15,73	0,30	3,84	0,39	0,43	0,23	0,23	0,17
CR	6234	6641	1,05	3,20	0,79	1,06	2,43	4,23	1,93	2,87	0,42	1,41	1,43	5,21	0,30	0,76
PR	3370	3178	0,03	0,43	0,00	0,00	0,00	0,33	0,00	0,27	0,00	0,00	0,67	7,66	0,00	0,00
PR	7442	7688	10,30	40,73	14,50	37,70	0,07	3,00	0,00	0,03	0,00	8,63	2,40	0,63	0,10	0,20
CR	4622	5001	25,33	34,20	11,03	7,39	4,70	3,57	14,57	18,61	4,73	4,87	2,53	4,15	1,01	1,52
PR	6429	5660	17,33	5,98	17,83	1,58	8,45	1,07	23,96	2,33	11,61	2,77	0,77	1,77	0,52	0,22
PD	-	3758	-	0,00	-	0,00	-	0,00	-	0,00		0,00	-	0,00	-	0,00
PD	4379	4131	0,90	0,70	0,10	0,60	0,17	0,23	0,03	0,03	0,00	0,00	0,20	0,80	0,05	0,60
PD	5559	5342	8,60	1,05	11,63	1,27	0,87	1,60	0,00	0,00	0,00	0,00	2,10	0,17	2,03	0,17
PD	3324	4748	4,39	0,98	2,80	2,60	15,40	9,39	7,07	6,49	0,86	0,54	1,53	0,73	0,37	0,30
PD	7149	6999	25,12	13,47	28,47	21,93	17,97	7,35	5,80	2,65	11,79	4,60	0,27	0,13	0,03	0,07
PD	4178	4513	5,20	7,23	2,37	1,10	1,80	1,17	10,33	3,87	1,29	0,83	0,45	0,13	0,12	0,07
PD	6552	8078	0,00	0,17	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
PD	6834	6977	31,99	3,09	3,69	1,03	32,37	3,22	8,69	0,28	1,32	0,02	24,42	3,70	2,26	0,46
PD	6422	6715	5,97	0,67	0,30	0,26	3,22	1,48	5,33	0,42	0,13	0,02	3,08	1,48	0,14	0,08
PD	4835	4637	2,30	0,87	0,99	0,55	0,47	0,48	0,42	0,51	0,00	0,00	1,80	1,33	0,51	0,43

Supplementary Table S6: Clinical variable values comparing patients with PD-1 blockade to those with PD-L1 blockade.

	All patients ***N*****=114 (100%)**	**PD-1 inhibitor therapy** **n=57 (100%)**	**PD-L1 inhibitor therapy** **n=57 (100%)**
Patient characteristics			
Sex			
male	82 (72%)	40 (70%)	42 (74%)
female	32 (28%)	17 (30%)	15 (26%)
Age			
< 70	40 (35%)	19 (33%)	21 (37%)
≥ 70	74 (65%)	38 (67%)	36 (63%)
Overall performance status (ECOG)			
0	64 (56%)	32 (56%)	32 (56%)
≥1	49 (43%)	25 (44%)	24 (42%)
Immunosuppression			
no	92 (81%)	44 (77%)	48 (84%)
yes	22 (19%)	13 (23%)	9 (16%)
LDH			
normal	43 (38%)	21 (37%)	22 (39%)
elevated	67 (59%)	35 (61%)	32 (56%)
CRP			
normal	30 (26%)	11 (19%)	19 (33%)
elevated	55 (48%)	27 (47%)	28 (49%)
NLR			
< 4	54 (47%)	22 (39%)	32 (56%)
≥ 4	35 (31%)	17 (30%)	18 (32%)
Tumor characteristics			
Localization of primary			
head and neck	24 (21%)	11 (19%)	13 (23%)
extremities	44 (39%)	17 (30%)	27 (47%)
trunk	19 (17%)	7 (12%)	12 (21%)
unknown	15 (13%)	10 (18%)	5 (9%)
Metastatic stage (AJCC)			
M0	17 (15%)	3 (5%)	14 (25%)
M1a	36 (32%)	21 (37%)	15 (26%)
M1b/M1c	61 (53%)	33 (58%)	28 (49%)
Organs onvolved			
1	51 (45%)	26 (46%)	25 (44%)
> 1	63 (55%)	31 (54%)	32 (56%)
MCPyV status (tumor)			
negative	10 (9%)	5 (9%)	5 (9%)
positive	32 (28%)	24 (42%)	8 (14%)
n.d.	72 (63%)	28(49%)	44 (77%)
PD-L1 (tumor)			
negative	17 (15%)	10 (17%)	7 (12%)
positive	21 (18%)	13 (23%)	8 (14%)
n.d.	76 (67%)	34 (60%)	42 (74%)
Therapeutic interventions			
Previous radiotherapy			
no	55 (48%)	24 (42%)	31 (54%)
yes	59 (52%)	33 (58%)	26 (46%)
Previous chemotherapy			
no	83 (73%)	38 (67%)	45 (79%)
yes	31 (27%)	19 (33%)	12 (21%)
Therapy response			
CR	24 (21%)	13 (23%)	11 (19%)
PR	30 (26%)	19 (33%)	11 (19%)
SD	20 (18%)	9 (16%)	11 (19%)
PD	40 (35%)	16 (28%)	24 (43%)

Abbreviations: AJCC, American Joint Committee on Cancer; ECOG, Eastern Cooperative Oncology Group; LDH, lactate dehydrogenase; MCPyV, Merkel cell polyomavirus; NLR, neutrophil to lymphocyte ratio.

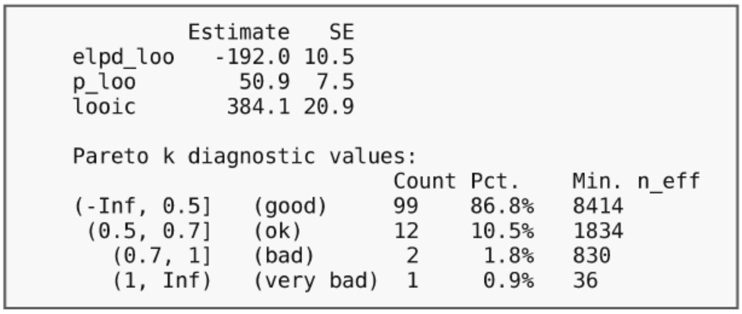
Supplementary Figure 1



Scatter plots comparing the actually observed data with 30 imputed data sets of clinical parameters with missing values in the Bayesian ordinal regression model. CRP – C-reactive protein; LDH – Lactate dehydrogenase; MCPyV – Merkel cell polyomavirus.

Supplementary Figure 2

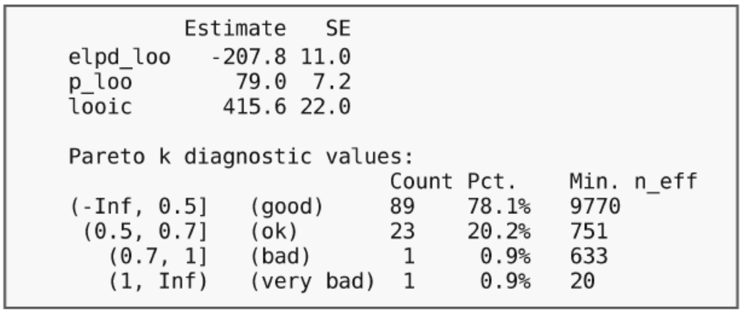
LOO sequential model with proportional odds assumption (fit_1):



```
          Estimate   SE
elpd_loo   -192.0 10.5
p_loo        50.9  7.5
looic       384.1 20.9

Pareto k diagnostic values:
                         Count Pct.    Min. n_eff
(-Inf, 0.5]   (good)     99    86.8%   8414
 (0.5, 0.7]   (ok)       12    10.5%   1834
   (0.7, 1]   (bad)       2     1.8%   830
   (1, Inf)   (very bad)  1     0.9%   36
```

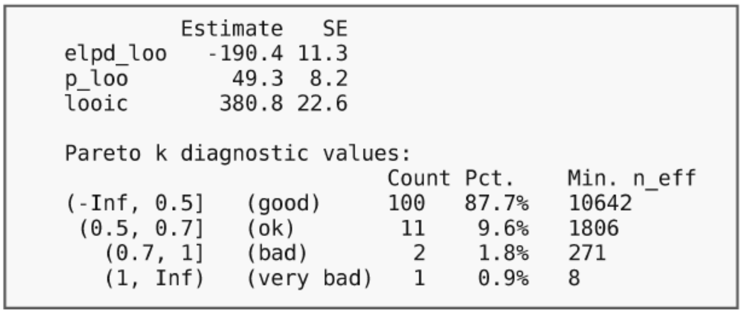
LOO sequential model without proportional odds assumption, category-specific parameters for all predictors (fit_2):



```
          Estimate   SE
elpd_loo   -207.8 11.0
p_loo        79.0  7.2
looic       415.6 22.0

Pareto k diagnostic values:
                         Count Pct.    Min. n_eff
(-Inf, 0.5]   (good)     89    78.1%   9770
 (0.5, 0.7]   (ok)       23    20.2%   751
   (0.7, 1]   (bad)       1     0.9%   633
   (1, Inf)   (very bad)  1     0.9%   20
```

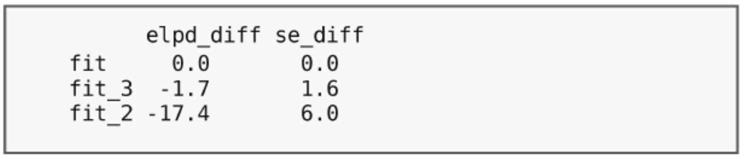
LOO sequential model with proportional odds assumption, category-specific parameters for most relevant predictors (immunosuppression, organs_greater_1) (fit_3):



```
          Estimate   SE
elpd_loo   -190.4 11.3
p_loo        49.3  8.2
looic       380.8 22.6

Pareto k diagnostic values:
                         Count Pct.    Min. n_eff
(-Inf, 0.5]   (good)     100   87.7%   10642
 (0.5, 0.7]   (ok)        11    9.6%   1806
   (0.7, 1]   (bad)        2    1.8%   271
   (1, Inf)   (very bad)   1    0.9%   8
```

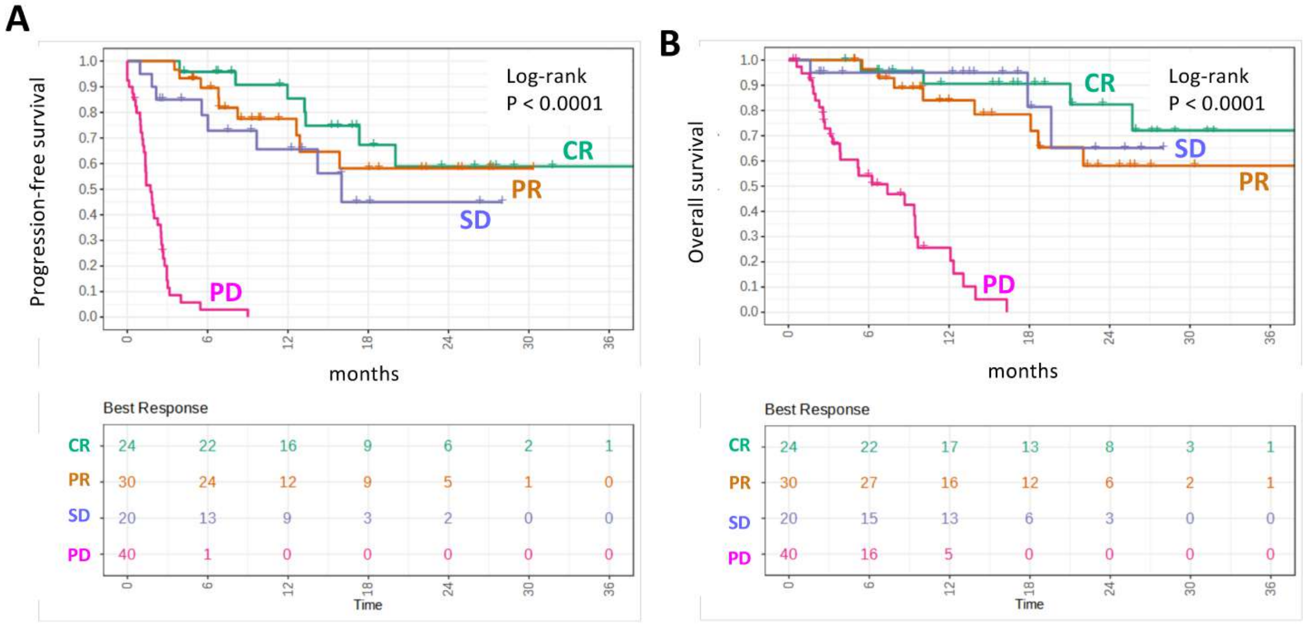
LOO comparisons (fit 1 vs. fit 2 vs. fit 3):



```
      elpd_diff se_diff
fit     0.0       0.0
fit_3  -1.7       1.6
fit_2 -17.4       6.0
```

Leave-one-out cross-validation (LOO) demonstrates that the choosen model has a similar ELPD (expected log-predictive density, a measure of its ability to generalize to unseen data) as a sequential model without category-specific effects, meaning that including category-specific effects does not improve model performance and the proportional odds assumption does not have a strong effect on the model conclusions.

Supplementary Figure 3



A

Best Response	0	6	12	18	24	30	36
CR	24	22	16	9	6	2	1
PR	30	24	12	9	5	1	0
SD	20	13	9	3	2	0	0
PD	40	1	0	0	0	0	0

Time

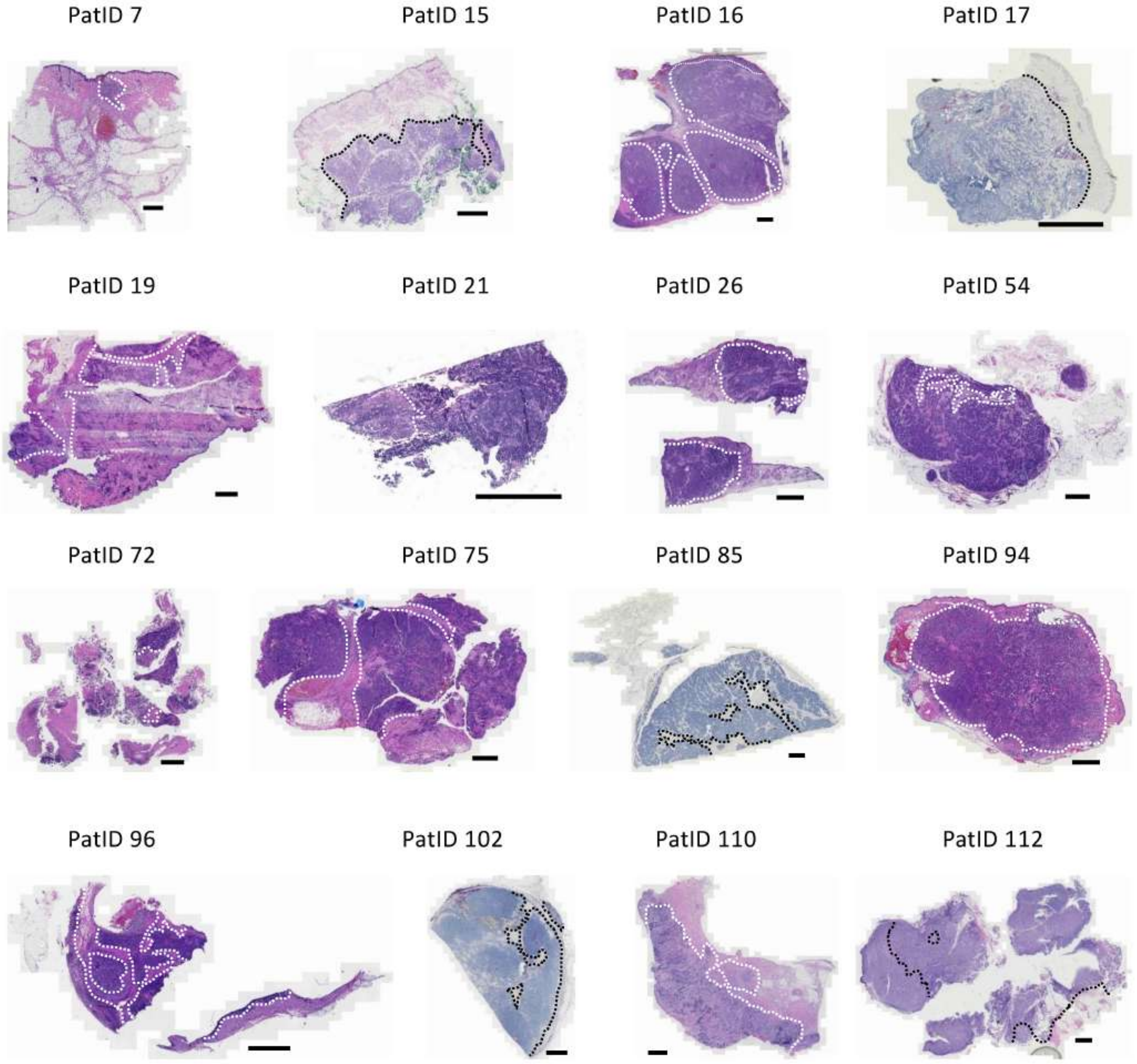
B

Best Response	0	6	12	18	24	30	36
CR	24	22	17	13	8	3	1
PR	30	27	16	12	6	2	1
SD	20	15	13	6	3	0	0
PD	40	16	5	0	0	0	0

Time

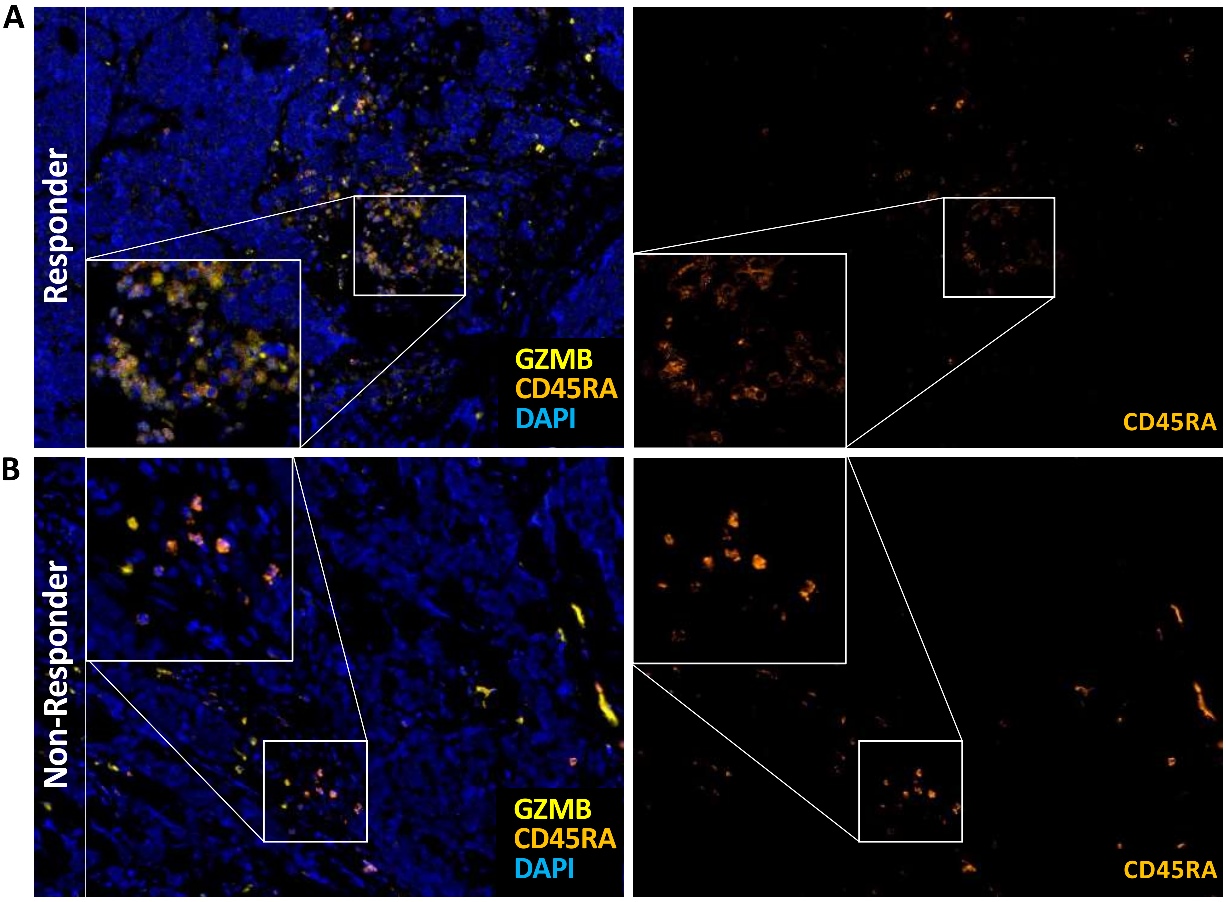
Survival characteristics of n = 114 advanced MM patients upon ICI therapy. (A, B) Kaplan Meier plots depicting (A) progression free (PFS) and (B) overall survival (OS) by BOR (CR, green, n=24; PR, orange, n = 30; SD, purple, n=20; PD, magenta, n=40) after therapy start. Patients after 36 months are all censored. CR, complete response; PR, partial response; SD, stable disease; PD, progressive disease.

Supplementary Figure 4



Whole slide scans of H&E stained tissue samples used for mIF based analysis of the immune infiltrate. Dotted lines (black/white) indicate separation between juxtra-tumoral and intra-tumoral areas (scale bar 2mm). Out of 114 samples, 21were subjected to mIF analysis. From these, 16 samples were already H&E stained for diagnostic reasons and were used here to depict the intra- and juxtra-tumoral regions. Slides were scanned with the Zeiss AxioScanZ.1 with 10x magnification.

Supplementary Figure 5



Excerpt of the multiplexed immunofluorescence staining of pre-treatment tumor tissue from a representative patient with disease control (A) and disease progression (B). Left: Co-expression of CD45RA (orange) and GZMB (yellow) displays high abundance of effector T cells in disease control. Right: Overall elevated presence of CD45RA positive cells in therapy responding patients.(20x magnification, DAPI counterstain)

Applying a cumulative ordinal regression model to infer possible biomarkers associated with response to PD-1/PD-L1 inhibition in Merkel cell carcinoma

Summary
Merkel cell carcinoma a is type of neuroendocrine skin cancer that in some cases can be treated with anti-PD-1/PD-L1 antibodies that act as immune checkpoint inhibitors and therefore enhance immune response against tumor cells. In an effort to identify biomarkers that distinguish treatment responders from non-responders, data of 114 patients had been collected and analyzed using a cumulative ordinal regression model. Conditioned on the model and the observed data, there is moderate statistical evidence that absence of immunosuppression, usage of anti-PD-1 antibodies (as opposed to anti-PD-L1 antibodies), and limited spread of the main tumor are associated with a higher probability of responding to the treatment.

Contents

1. Dataset

The dataset had been collected from 4 patients with Merkel cell carcinoma. All patients were either treated with anti-PD-1 or anti-PD-L1 antibodies and treatment response was classified into progressive disease (PD), stable disease (SD), partial response (PR) and complete response (CR). In total, there are 19 different predictors:

- gender (categorical: male, female)
- primary localisation
  (categorical: head + neck, occult, extremities, trunk)
- immunosuppression (binary)
- tumor PD-L1 expression (binary)
- MCPyV+ status (binary)
- prior chemotherapy (binary)
- prior radiotherapy (binary)
- checkpoint inhibition (categorical: PD-1, PD-L1)
- metastasic stage (categorical: M0, M1a, M1b/M1c)
- ≥ 2 organs involved (binary)
- elevated LDH levels (binary)
- elevated CRP levels (binary)
- neutrophil count at therapy start (numeric)
- lymphocyte count at therapy start (numeric)
- neutrophil/lymphocyte ratio (NLR) ≥ 4 (binary)
- ECOG performance status ≥ 1 (binary)
- age ≥ 70 years (binary)
- year of therapy start (ordered categorical)
- participation in a clinical trial (binary)

2. Introduction

To analyze these data, we fit a Bayesian model. This has several advantages:

A unique feature of Bayesian statistics is that it allows to describe model parameters with probability distributions. This means that instead of point estimates (with more or less reliable standard deviations) we obtain a distribution of parameter values that is consistent with the observed data. In this way, it is possible to quantify the uncertainty of the estimates.

Additionally, Bayesian statistics provides an accessible way to test models: By comparing data generated under the model's assumptions to the actually observed data, it is possible to identify important aspects of the dataset that the model fails to capture, and subsequently improve the model until it is consistent with the observed data.

The dataset also contains missing values, and Bayesian statistics allows to incorporate data points where some of the inputs are missing in such a way that the uncertainty of the missing values directly translates into uncertainty of the estimates.

3. Model description

The treatment response (progressive disease, stable disease, partial response or complete response) is on an ordinal scale, which means that the different levels have an inherent order, e.g a partial response is clearly better than stable disease.

Unfortunately, statistical models that are applied to this kind of data often do not adequately account for this order: In practice, a common approach is to encode the categories as increasing integer values (1, 2, 3, ...) and to apply a linear regression model. While this preserves the order, it assumes equal distances between the outcomes, which means that the distance between progressive- and stable disease is identical to the distance between partial response and complete response. Other common approaches involve multinomial models (which ignore the order) or fitting logistic regression models after arbitrarily binarizing the response.

All these approaches share the common problem that they do not make efficient use of the available data and might lead to over- or underestimated effect sizes [1]. For instance when binning the data into two categories (responder and non-responder), patients with a partial response have the same influence on the overall estimates as patients with a complete response. The grouping is often also arbitrary, for example in terms of survival time, our data shows that patients with stable disease are actually more similar to partial- and complete responders than to patients with progressive disease (data not shown).

3.1 Cumulative ordinal regression model

To circumvent the aforementioned issues, we apply a cumulative ordinal regression model, which adequately takes the order between the categories into account.

The main idea of the cumulative regression model is to regard the tendency of a patient to respond to the treatment as a latent (= unobserved) variable that is determined by the patient characteristics. By convention, this latent variable is usually assumed to be distributed according to a logistic distribution with a scale of 1 and a mean that is determined by the predictor values for that patient.

The logistic distribution with a scale of 1 closely resembles a normal distribution with a standard deviation of 1.6. In fact, the exact choice of distribution does not matter in practice and other choices are possible, e.g. using a standard normal distribution instead would lead to a class of models called probit models (whereas the cumulative regression model using the logistic distribution is a generalization of the logistic regression model for ordered responses with more than two categories).

The probabilities of the four possible response categories are then determined by three thresholds that are also estimated.

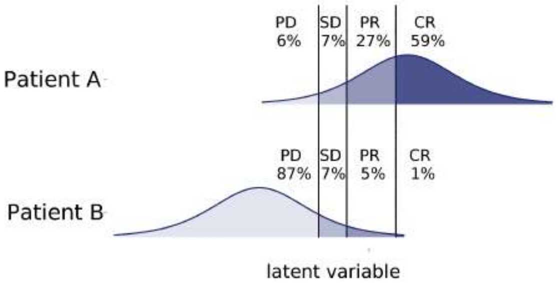


Figure 1. Distributions of the inferred latent variable for two patients in the dataset. PD: progressive disease, SD: stable disease, PR: partial response, CR: complete response.

Figure 1 shows the distribution of the latent variable for two patients in the dataset along with the 3 estimated thresholds (vertical lines). Patient A has predictor values that favor treatment response, whereas Patient B has predictor values that do not favor treatment response. The probability for each of the response categories is given by the area under the curve that is enclosed by the corresponding thresholds. Please note that, as we fit the model in a Bayesian fashion, neither the location of the latent variables nor the tresholds are fixed values (as shown in Fig. 1), but follow some distribution of values that are consistent with the observed data.

In formula notation, the model can be written as:

$$\begin{aligned}
\mu_i &= \beta_{\text{age}} \cdot x_{\text{i, age}} + \beta_{\text{LDH}} \cdot x_{\text{i, LDH}} \cdots \\
p(\text{PD})_i &= \int_{-\infty}^{\tau_1} \text{logistic\_distribution}(\mu_i, 1)dx \\
p(\text{SD})_i &= \int_{\tau_1}^{\tau_2} \text{logistic\_distribution}(\mu_i, 1)dx \\
p(\text{PR})_i &= \int_{\tau_2}^{\tau_3} \text{logistic\_distribution}(\mu_i, 1)dx \\
p(\text{CR})_i &= \int_{\tau_3}^{\infty} \text{logistic\_distribution}(\mu_i, 1)dx
\end{aligned} \tag{1}$$

,

where

- μ_i is the location of the latent variable for patient i,
- β_{age} is the β coefficient for the predictor age,
- $x_{\text{i, age}}$ is the indicator variable of patient i for age (in this case, 0 if patient i's age is ≥ 70 years, 1 otherwise),
- $p(\text{PD})_i$ is the probability of a progressive disease response
- τ_1, τ_2, τ_3 are the three estimated tresholds.

The β coefficients of the predictors are of main interest in this analysis, as they give information on whether or not a predictor is associated with a higher probability of responding to the treatment. A more comprehensive explanation of ordinal regression models that is also accessible without a background in statistics is given in [2].

3.2 Model fitting

Fitting the model to the dataset was done with the R software package 'brms' [3], which utilizes 'Stan' [4] in the background. Student t priors with 7 degrees of freedom and a standard deviation of 1 were chosen as weakly informative priors for the β coefficients. This is in line with the Stan prior choice recommendations [1]. The t distribution has a similar shape as the normal distribution, but with higher density in the tail areas. In this way, we rule out unreasonably large parameter values (e.g. anything larger than 10-15 for coefficients on the log-odds scale), but the model is still flexible enough to allow for values that might make sense. Model fit is performed numerically by Markov chain Monte Carlo. In total, 2000 samples from 4 different Markov chains were generated. We use the split-$\hat{\text{R}}$ diagnostic [5, 6] to identify possible Markov chain convergence issues. All parameters satisfied $\hat{\text{R}} < 1.01$, the effective sample size N_{eff} [7] exceeded 1000 in all cases.

4. Model results

Figure 2 shows marginal posterior distributions of the estimated β coefficients. Values larger than 0 denote that these predictors favor a response to the treatment, whereas values less than 0 favor treatment non-response. The width of the distribution gives an impression of the uncertainty of the estimate: A distribution tightly concentrated around some value means that the dataset allows for a precise estimate of that parameter, while a broader distribution means that the data is consistent with a wide range of parameter values.

Please note that all these estimates are conditioned on the model and the observed data, which means that they are not a statement about the general population of patients with Merkel cell carcinoma.

Most of estimates include 0, which means that the absence of association between that predictor and the treatment response is a reasonable explanation for the observed data. The widths of the distributions are also broad, so while no effect is a possible explanation, it could also be quite large.

Notable exceptions are the predictors immunosuppression and organs involved, where most of the probability mass is located at values less than 0 (denoting they are associated with a decreased probability of treatment response); and the use of an anti-PD-1 antibody, where most of the probability mass is located at values greater than 0 (denoting it is associated with an increased probability of treatment response), as compared to checkpoint inhibition with an anti-PD-L1 antibody.

4.1 Average Predictive Comparisons

As with logistic regression models, the β coefficients of cumulative ordinal regression models are in units of log-odds, which means that a value of 1 of the corresponding predictor increases the expected log-odds of the next higher response category by 1.

[1] https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations

This has the disadvantage that it is difficult to have an intuition about the effect size, i.e. whether an increase in the log-odds by 1 corresponds to a large, moderate or small change. To circumvent this issue, we show average predictive comparisons in addition to the regression coefficients.

Briefly, average predictive comparisons are calculated as the expected change in the response associated with a unit difference in one of the inputs. A technical description of average predictive comparisons is given in [8]. In this analysis, it allows us to reinterpret the regression coefficients (which are in units of log-odds) into a summary that is on the probability scale.

Figure 2 shows average predictive comparisons for each of the model inputs. They were calculated with respect to having at least a partial response to the treatment, e.g. for immunosuppression, values between -0% and -40% denote that comparing a patient with immunosuppression to an otherwise identical patient without immunosuppression, the patient with immunosuppression has (on average) a 0% to 40% lower probability of having at least a partial response to treatment.

As a result of the limited sample size, the uncertainty around the estimates is rather large. It also shows that just because a predictor includes 0%, it should not be confused with having no association with the treatment response, because the data is still consistent with large effect sizes in either direction.

5. Imputing missing values

The dataset contains missing values in some of the predictors. Standard practice is usually to delete them, either by row-wise exclusion (removing all samples that contain any missing value), or by removing the predictors that contain missing values.

With only 114 patients, removing all samples that contain missing values would mean to remove import information.

As a more sensible approach, multiple imputation with the R package mice [9] was used instead. In multiple imputation, several imputed versions of the dataset are created where the missing values are replaced with plausible values. The imputed datasets are identical for the non-missing entries, but differ in the imputed values. The uncertainty about the missing values is reflected in the degree of variation between the datasets.

To translate these different datasets into as single estimate, we simply fit the model independently on each dataset and combine the posterior samples of each model fit. In this way, the uncertainty of the missing values propagates directly into uncertainty of the estimates.

6. Model testing

A useful way to test Bayesian models is called posterior predictive check. In posterior predictive checks, the inferred parameter estimates are used to sample an arbitrary number of new datasets that are generated under the model's assumptions. By comparing these datasets to the actually observed dataset, it is possible to identify aspects of the data that the model fails to capture.

One possible way to perform a posterior predictive check for the model described here is to compare the observed proportion of the different treatment response categories to the proportion of treatment response categories expected under the model's assumptions. Figure 3 shows a histogram of the proportion of patients with progressive disease, stable disease, partial response and complete response in the generated datasets with the actually observed proportions highlighted in blue. The observed proportions lie directly in the center of what is expected by the model.

Another form of posterior predictive check focuses on individual predictions instead. For each patient, the expected probability that a given patient has at least a partial response to the antibody treatment is calculated. If the model produces reasonable estimates, we expect that patients with a higher estimated probability really do respond more frequently to the treatment than patients with a lower estimated probability.

Figure 4 shows a so-called calibration plot. All 114 patients were sorted according to their expected probability of having at least a partial response to the treatment and placed into 7 distinct bins. For each bin, the mean probability of having a partial response is plotted against the observed proportion of patients in that bin which at least partially respond to the treatment. As each uncertainty interval around the observed proportion touches the diagonal line, this type of posterior predictive check shows again no large discrepancies between expected and observed data.

In conclusion, the posterior predictive checks show that the clinical data is consistent with data expected under the model's assumptions.

Applying a cumulative ordinal regression model to infer possible biomarkers associated with response to PD-1/PD-L1 inhibition in Merkel cell carcinoma — 5/6

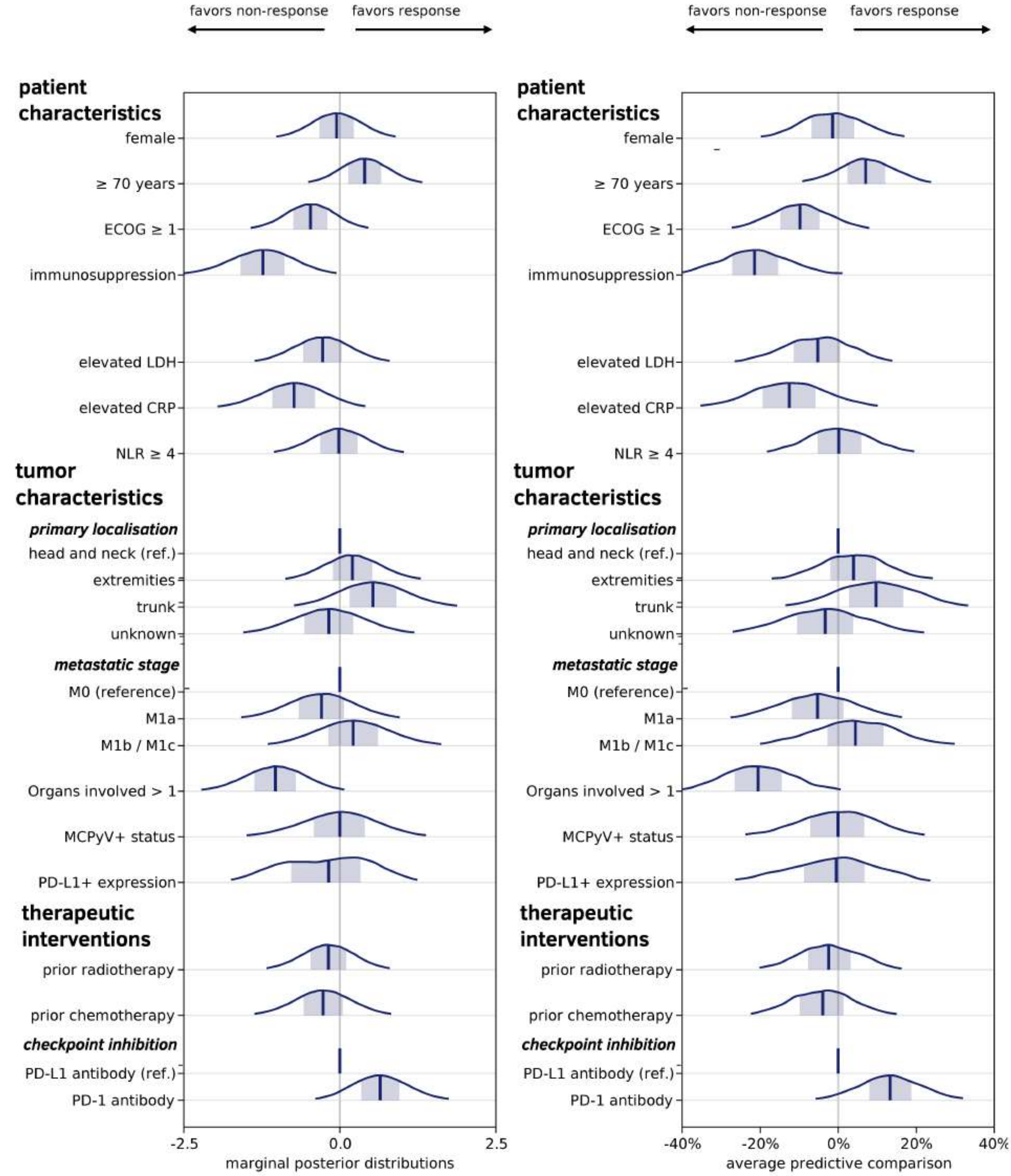


Figure 2. Marginal posterior distributions of β coefficients (left) and average predictive comparisons of the expected probability of having at least a partial response to the treatment (right). The regression coefficients of the full model have been projected onto a probability scale.

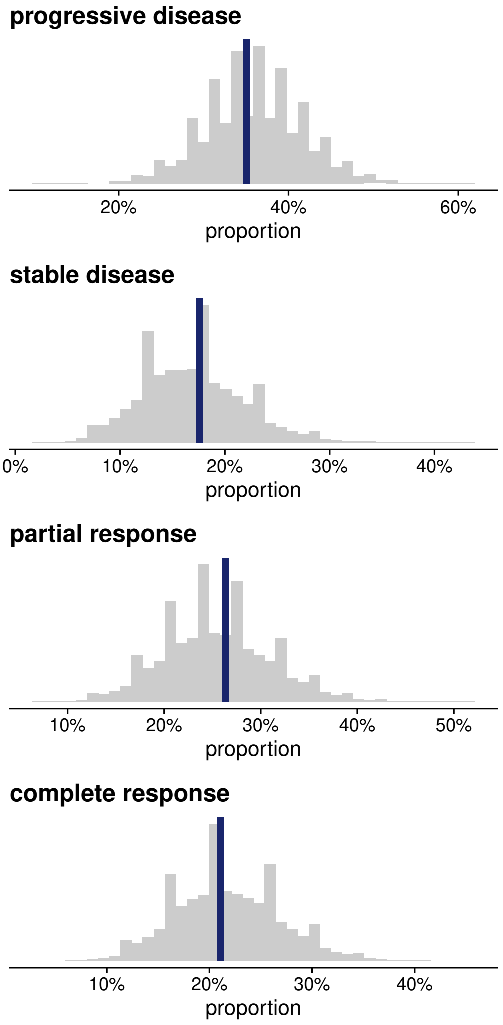


Figure 3. Expected proportion of each response category under the model (histogram) vs. observed proportion (blue line).

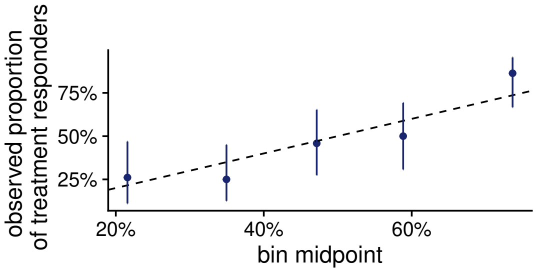


Figure 4. Calibration plot of proportion of patients with at least a partial response to the treatment.

References

[1] Torrin M. Liddell and John K. Kruschke. Analyzing ordinal data: Support for a Bayesian approach. *SSRN Electronic Journal*, 2015.

[2] Paul-Christian Bürkner and Matti Vuorre. Ordinal regression models in psychology: A tutorial. 2:77–101, 2019.

[3] Paul-Christian Bürkner. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1), 2017.

[4] Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 2017.

[5] Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, nov 1992.

[6] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

[7] Stan Development Team. *Stan Modeling Language Users Guide and Reference Manual, 2.23. https://mc-stan.org*, 2021.

[8] Andrew Gelman and Iain Pardoe. 2. Average Predictive Comparisons for models with nonlinearity, interactions, and variance components. *Sociological Methodology*, 37(1):23–51, aug 2007.

[9] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 2011.

3.3 Predominance of Central Memory T Cells with High T-Cell Receptor Repertoire Diversity is Associated with Response to PD-1/PD-L1 Inhibition in Merkel Cell Carcinoma

This section is based on the following publication:

Ivelina Spassova, Selma Ugurel, Patrick Terheyden, Antje Sucker, Jessica C Hassel, Cathrin Ritter, Linda Kubat, Daniel Habermann, Farnoush Farahpour, Mohammadkarim Saeedghalati, Lukas Pfeiffer, Rajiv Kumar, David Schrama, Daniel Hoffmann, Dirk Schadendorf, Jürgen C. Becker (2020).
Predominance of Central Memory T Cells with High T-Cell Receptor Repertoire Diversity is Associated with Response to PD-1/PD-L1 Inhibition in Merkel Cell Carcinoma

Predominance of Central Memory T Cells with High T-Cell Receptor Repertoire Diversity is Associated with Response to PD-1/PD-L1 Inhibition in Merkel Cell Carcinoma

Ivelina Spassova[1], Selma Ugurel[2], Patrick Terheyden[3], Antje Sucker[2], Jessica C. Hassel[4], Cathrin Ritter[1], Linda Kubat[1], Daniel Habermann[5], Farnoush Farahpour[5], Mohammadkarim Saeedghalati[5], Lukas Peiffer[1,6], Rajiv Kumar[6,7], David Schrama[8], Daniel Hoffmann[5], Dirk Schadendorf[2], and Jürgen C. Becker[1,2,6]

ABSTRACT

Purpose: Merkel cell carcinoma (MCC) is an aggressive neuroendocrine skin cancer, which can be effectively controlled by immunotherapy with PD-1/PD-L1 checkpoint inhibitors. However, a significant proportion of patients are characterized by primary therapy resistance. Predictive biomarkers for response to immunotherapy are lacking.

Experimental Design: We applied Bayesian inference analyses on 41 patients with MCC testing various clinical and biomolecular characteristics to predict treatment response. Further, we performed a comprehensive analysis of tumor tissue–based immunologic parameters including multiplexed immunofluorescence for T-cell activation and differentiation markers, expression of immune-related genes and T-cell receptor (TCR) repertoire analyses in 18 patients, seven objective responders, and 11 nonresponders.

Results: Bayesian inference analyses demonstrated that among currently discussed biomarkers only unimpaired overall performance status and absence of immunosuppression were associated with response to therapy. However, in responders, a predominance of central memory T cells and expression of genes associated with lymphocyte attraction and activation was evident. In addition, TCR repertoire usage of tumor-infiltrating lymphocytes (TILs) demonstrated low T-cell clonality, but high TCR diversity in responding patients. In nonresponders, terminally differentiated effector T cells with a constrained TCR repertoire prevailed. Sequential analyses of tumor tissue obtained during immunotherapy revealed a more pronounced and diverse clonal expansion of TILs in responders indicating an impaired proliferative capacity among TILs of nonresponders upon checkpoint blockade.

Conclusions: Our explorative study identified new tumor tissue–based molecular characteristics associated with response to anti–PD-1/PD-L1 therapy in MCC. These observations warrant further investigations in larger patient cohorts to confirm their potential value as predictive markers.

Introduction

Merkel cell carcinoma (MCC) is a highly aggressive neuroendocrine skin cancer, which occurs predominantly in the elderly, fair-skinned population. The mortality rate after primary diagnosis is reported to range from 33% to 46% (1). Known risk factors for MCC include chronic UV exposure and any type of immunosuppression (2). Moreover, an association with the Merkel cell polyomavirus (MCPyV) has been established predominantly for cases occurring in the northern hemisphere (3). Key feature of MCC is its high immunogenicity based on either its virus- or UV-associated carcinogenesis, causing either presentation of MCPyV-derived epitopes (4) or neo-epitopes by UV-associated mutations (5). Indeed, a high therapeutic activity of immune checkpoint inhibitors (CPIs) resulting in durable objective responses (ORs) in about 50% of patients has been observed (6, 7). Predictive biomarkers of therapy outcome to differentiate responders from nonresponders at treatment baseline have not been established. On the basis of the experience in melanoma, clinical features like overall performance status, metastatic stage, and previous therapies, as well as blood-derived parameters, for example, neutrophil-to-lymphocyte ratio, elevated lactate dehydrogenase (LDH) and C-reactive protein, are currently discussed as possible predictive markers (8, 9). However, their predictive power could so far not been indisputably confirmed (10). Similarly, potential predictive biomarkers determined on tumor tissue such as PD-L1 expression or MCPyV status could not be confirmed either (6, 7).

This study provides an extensive workup of both, clinical as well as immunologic and molecular features of 41 patients with MCC treated with CPIs for advanced disease in the real-world setting. The latter were based on T-cell receptor (TCR) repertoire usage, as well as gene and protein expression determined in a subgroup of patients on formalin-fixed, paraffin-embedded (FFPE) tumor tissue obtained at

[1]Translational Skin Cancer Research, German Consortium for Translational Cancer Research (Deutsches Konsortium für Translationale Krebsforschung; DKTK), Essen, Germany. [2]Department of Dermatology, University Hospital of Essen, Essen, Germany. [3]Department of Dermatology, University Hospital of Lübeck, Lübeck, Germany. [4]Department of Dermatology, University Hospital of Heidelberg, Germany. [5]Bioinformatics and Computational Biophysics, University of Duisburg-Essen, Essen, Germany. [6]German Cancer Research Center (Deutsches Krebsforschungs Zentrum, DKFZ), Heidelberg, Germany. [7]Division of Molecular Genetic Epidemiology, Heidelberg, Germany. [8]Department of Dermatology, University Hospital of Würzburg, Germany.

Note: Supplementary data for this article are available at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org/).

I. Spassova and S. Ugurel contributed equally to this article.

Corresponding Author: Jürgen C. Becker, German Cancer Research Center, Universitaetsstr. 1, S05 T05 B, Essen 45141, Germany. Phone: 4920-1183-6727; Fax: 4920-1183-6945; E-mail: j.becker@dkfz.de

Clin Cancer Res 2020;26:2257–67

doi: 10.1158/1078-0432.CCR-19-2244

Translational Relevance

Immunotherapy of advanced Merkel cell carcinoma by anti–PD-1/PD-L1 antibodies has greatly improved the prognosis of this highly aggressive neuroendocrine skin cancer. Unfortunately, about half of the patients have no durable benefit from immune checkpoint blockade. Thus, reliable predictive biomarkers are needed. Here, we report that tumor-infiltrating lymphocytes with a central memory phenotype and a diverse T-cell receptor repertoire correlate with a favorable response to immunotherapy.

baseline before start of anti–PD-1/PD-L1 therapy. These unbiased multidimensional analyses were performed to improve our understanding of MCC immunology and to identify new candidates for predictive biomarkers. To further extend this aim, some of the patients' sequential samples obtained before and under therapy were analyzed.

Our results demonstrate that among previously discussed clinical and molecular biomarkers, only an unpaired performance status and the absence of immune suppression were strongly associated with a favorable clinical response to immunotherapy. More important, however, not the mere density of the immune infiltrate, but rather its functional properties correlated with response to CPI treatment. Specifically, tumor-infiltrating lymphocytes (TILs) with a predominance of central memory T cells and a diverse TCR repertoire were associated with a favorable treatment outcome.

Materials and Methods

Patients and samples

Patients treated at the Departments of Dermatology, University Hospitals of Essen, Heidelberg and Lübeck, were retrospectively identified for this biomarker study according to the following selection criteria: diagnosis of MCC confirmed by histopathology, metastatic disease not amendable to surgery, and systemic therapy with anti–PD-1/PD-L1 antibodies. Treatment response to anti–PD-1/PD-L1 CPI was categorized as best overall response according to RECIST v1.1 (11). The study was approved by the ethics committee of the University Duisburg-Essen (11-4715; 17-7538-BO) and was conducted in accordance with the Declaration of Helsinki—Ethical Principles for Medical Research Involving Human Subjects. Informed written consent was obtained from each subject.

Detection of MCPyV DNA

Detection of MCPyV DNA was performed as described previously (12). Briefly, DNA from FFPE tissue samples was isolated by AllPrep DNA/RNA FFPE Kit (Qiagen) according to the manufacturer's instructions. Presence of MCPyV DNA was determined by TaqMan Real-Time qPCR using large T-antigen–specific primers and TaqMan probe: forward primer: CCA AAC CAA AGA ATA AAG CAC TGA; reverse primer: TCG CCA GCA TTG TAG TCT AAA AAC; and probe: FAM-AGC AAA AAC ACT CTC CCC ACG TCA GAC AG-BHQ1. The PCR reaction had a final volume of 10 μL, consisting of DNA (10 ng), primers and probe (5 μmol/L each), and reaction buffer (LuminoCt ReadyMix, Sigma-Aldrich). Annealing was performed at 60°C for 15 seconds. CFX Manager (Bio-Rad) was used for data analysis.

IHC quantification of PD-L1 expression

PD-L1 expression was assessed in FFPE tumor tissue sections with the use of a rabbit monoclonal anti-human PD-L1 antibody (clone 28-8) and an analytically validated automated IHC assay (PD-L1 IHC 28-8 pharmDx for Autostainer Link 48; Dako), as described previously (13). PD-L1 positivity was defined as at least 1% of living tumor cells showing specific cell surface staining of any intensity in a section containing at least 100 evaluable tumor cells. Positive staining of tumor-infiltrating inflammatory cells or other cells of the tumor stroma was excluded from evaluation.

TCR beta-chain clonotype mapping

The TCR repertoire usage of TILs was determined by amplifying the highly variable CDR3 of the different TCR beta-chain (TCRB) families by a multiplexed PCR and subsequent high-throughput sequencing using the immunoSEQ hsTCRB Kit (Adaptive Biotechnologies, catalog No. ISK10101). After DNA was isolated from FFPE tumor tissue in a first PCR reaction, all recombined TCRB CDR3 sequences were amplified by a mix of V- and J-gene primers. After labeling the obtained amplicons in a second PCR amplification, the resulting library was sequenced on an Illumina MiSeq using the MiSeq Reagent Kit v3 (Illumina, catalog No. MS-102-3001).

Multiplex immunofluorescence staining

Multiplex immunofluorescence staining of FFPE tumor tissue was performed using the OpalTM Chemistry (PerkinElmer, catalog No. OP7TL4001KT) with two panels of antibodies, that is, against CD4, CD8, CD45RA, CD45RO, and CK20 (panel 1), or CD27, CD45RA, CD45RO, and Synaptophysin (panel 2). Briefly, after deparaffinization and fixation, 3-μm tumor sections were processed with retrieval buffers for 15 minutes in an inverter microwave oven. Thereafter, sections were incubated with the antibody diluent for 10 minutes at room temperature, followed by incubation with the primary antibody for 30 minutes. After applying Opal Polymer HRP secondary antibody solution for 10 minutes, antibodies were removed by microwave treatment before another round of staining was performed. The antibodies and retrieval buffers used are described in detail in Supplementary Table S1. Visualization of the different fluorophores was achieved on the Mantra Quantitative Pathology Imaging System (PerkinElmer).

Gene expression analysis

mRNA from FFPE tissue samples was isolated using the AllPrep DNA/RNA FFPE Kit (Qiagen, catalog No. 80234) according to the manufacturer's instructions. Gene expression was quantified using the HuV1 Cancer Immune Panel (NanoString Technologies, catalog No. XT-CSO-HIP1-12). After 100 ng of mRNA was used for the hybridization reaction at 65° C for 24 hours, the complex was further processed in the nCounter Prepstation for immobilization to the cartridge, which was processed in the nCounter Digital Analyzer.

Statistical and bioinformatic analyses

A Bayesian logistic regression model was applied for predicting PD-1/PD-L1 blockade treatment response. It was built on a dataset consisting of 41 patients and 15 clinical parameters (Supplementary Table S2). The observed data were described in a probabilistic manner, also known as likelihood function. The probability that a given sample belongs to one of the two possible outcomes, responding to CPI therapy or not, was computed via the following formula: $\theta_j = logistic(\beta_0 + \beta_{parameterX} \cdot x_{j,parameterX} \ldots)$; in which θ_j is the

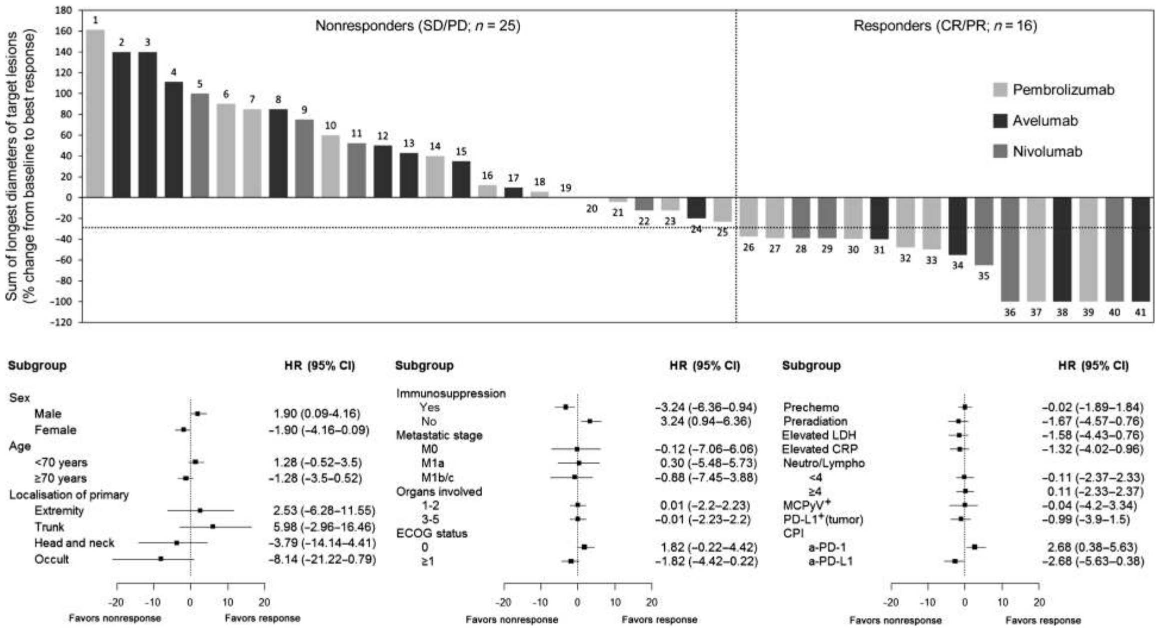


Figure 1.

Best overall response to anti–PD-1/PD-L1 therapy in correlation to baseline characteristics. Waterfall plot depicting best tumor response upon anti–PD-1/PD-L1 therapy as change in the sum of the longest diameters of target lesions from baseline to best response. Data for $n = 41$ patients with each bar, color-coded by therapeutic antibody, representing an individual patient are depicted. Pointed lines discriminate responders (CR and PR) from nonresponders (SD and PD) and the 30% decrease of the tumor volume classifying for OR. Clinical parameters at baseline and their correlation to therapy outcome are visualized by forest plots and HRs with 95% credibility intervals (CIs) calculated by Bayesian logistic regression model.

probability of responding to the treatment for patient j, β_0 is the base frequency of responders in the dataset, $\beta_{parameterX}$ is the effect of parameter X on the treatment response, and $x_{j,prameterX}$ is 1 if for patient j parameter X is applicable and 0 otherwise (14). In addition, HRs for each clinical parameter were calculated as mean of the credibility (posterior) interval. Fitting the model to the dataset was done with the R software package "brms," which utilizes "Stan" in the background. Missing values were approached with multiple imputation with the R package "mice" (15–17). For additional details, please see Supplementary Materials and Methods.

Normalization of the rank-abundance-distribution (RAD) of T-cell clonality was done by MaxRank normalization as implemented in R package RADanalysis (18). Here, the MaxRank is the minimum dimension of rank abundance vectors for all tested samples (18). The plotted normalized RADs are the results of 50-fold averaging. The distribution of T-cell numbers over T-cell clones was quantified by Pielou's Evenness Index J = H/H_{max}, with Shannon entropy H, and theoretically possible maximum Shannon entropy H_{max}. J ranges between 0 and 1, where 1 represents completely even distribution of T cells over clones (19). Observed richness indicates the number of unique T-cell clones, and the Chao index iChao1 is used as an estimator of TCR clone richness for rare clones (20). Simpson Diversity Index (also known as, Simpson D) represents the probability that two T cells taken at random from a specimen represent the same clone (19).

We implemented the Grouping of Lymphocyte Interactions by Paratope Hotspots (GLIPH) analysis method (21) to uncover TCR antigen specificities shared between clones and patients: CDR3 sequences of TCR clonotypes were clustered according to their local and global similarities; global similarity was assumed if only one amino acid was exchanged; and local similarity was assumed if specific motifs of three amino acids in the CDR3 region were more frequently present than in the reference database. The software R v3.4.1 and nSolver3 (NanoString Technologies) were used for gene expression data analysis. For P value adjustment the Benjamini–Hochberg algorithm was used. The respective correlation coefficients were calculated by the method of Pearson. Gene ontology (GO) analysis was performed using the online platform Metascape (http://metascape.org; ref. 22), where differentially expressed genes are assigned to a set of predefined terms (Kyoto Encyclopedia of Genes and Genomes Pathway, GO Biological Processes, Reactome Gene Sets, Canonical Pathways, and CORUM). For estimation of term similarity, the agreement calculator Cohen kappa coefficient was deployed; a kappa value > 0.3 was set as a threshold for selecting the terms for clustering.

For statistical testing of the TCR repertoire characteristics, gene expression analysis, and central memory T (T_{CM}) cells abundance in MCC tumor tissue, P values were determined using the unpaired two-tailed Student t test, calculated in GraphPad Prism 5.

Results

Overall performance status and absence of immunosuppression predict response to CPI

Forty-one patients treated with PD-1/PD-L1–blocking antibodies for advanced MCC were identified in three clinical centers. To determine the impact of the clinical and standard immunologic parameters on therapy response, we performed chart review. The extracted parameters were correlated by Bayesian inference with

Table 1. Baseline characteristics of patients included in the immunologic work-up.

	All patients *N* = 18 (100%)	Responders (CR/PR) *n* = 7 (100%)	Nonresponders (SD/PD) *n* = 11 (100%)
Gender			
Male	11 (61%)	6 (86%)	5 (45%)
Female	7 (39%)	1 (14%)	6 (55%)
Age			
≤70 years	10 (56%)	5 (71%)	5 (45%)
>70 years	8 (44%)	2 (29%)	6 (55%)
Localization of primary			
Extremities	8 (44%)	4 (57%)	4 (36%)
Head and neck	5 (28%)	1 (14%)	4 (36%)
Trunk	1 (6%)	1 (14%)	0 (0%)
Unknown primary	4 (22%)	1 (14%)	3 (27%)
Metastatic stage (AJCC)			
Skin, soft tissue, LN (M0)	3 (17%)	1 (14%)	2 (18%)
Skin, soft tissue, LN (M1a)	8 (44%)	2 (29%)	6 (55%)
Lung (M1b)	0 (0%)	0 (0%)	0 (0%)
Other organs (M1c)	7 (39%)	4 (57%)	3 (27%)
Organs involved	15 (83%)	5 (71%)	10 (91%)
1–2	3 (17%)	2 (29%)	1 (9%)
3–5			
Overall performance status (ECOG)			
0	8 (44%)	3 (43%)	5 (45%)
≥1	10 (56%)	4 (57%)	6 (55%)
Previous chemotherapy			
Yes	11 (61%)	4 (57%)	7 (64%)
No	7 (39%)	3 (43%)	4 (36%)
Previous radiotherapy			
Yes	10 (56%)	4 (57%)	6 (55%)
No	8 (44%)	3 (43%)	5 (45%)
LDH (blood)			
≤ULN	6 (33%)	3 (43%)	3 (27%)
>ULN	12 (67%)	4 (57%)	8 (73%)
MCPyV status (tumor)			
Positive	16 (89%)	6 (86%)	10 (91%)
Negative	2 (11%)	1 (14%)	1 (9%)
PD-L1 (tumor)			
Positive (≥1%)	8 (44%)	2 (29%)	6 (55%)
Negative (<1%)	10 (56%)	5 (71%)	5 (45%)
Not specified	0 (0%)	0 (0%)	0 (0%)
PD-1/PD-L1 inhibitor therapy			
Avelumab	4 (22%)	2 (29%)	2 (18%)
Nivolumab	6 (33%)	3 (43%)	3 (27%)
Pembrolizumab	8 (44%)	2 (29%)	6 (55%)
Best overall response to anti-PD-1/PD-L1			
CR	2 (11%)	2 (29%)	0 (0%)
PR	5 (28%)	5 (71%)	0 (0%)
SD	4 (22%)	0 (0%)	4 (36%)
PD	7 (39%)	0 (0%)	7 (64%)

Abbreviations: AJCC, American Joint Committee on Cancer; ULN, upper limit of normal.

clinical response to therapy. Patients showing an OR to anti–PD-1/PD-L1 therapy, that is, complete (CR) or partial response (PR), were classified as responders, and patients with a stable (SD) or progressive disease (PD) as nonresponders. At database closure (November 2018), the median follow-up time after onset of checkpoint inhibition was 13.5 months. Detailed patient characteristics are provided in Supplementary Table S2.

Bayesian inference demonstrated that only an unimpaired overall performance status [Eastern Cooperative Oncology Group (ECOG) = 0] is associated with a positive response to CPI treatment, whereas any form of immunosuppression was identified as a negative predictor (**Fig. 1**). However, even for the unimpaired performance status, the 95% credibility interval overlaps with the region of zero effect. For all other currently discussed predictive biomarkers for response to CPI such as metastatic stage, neutrophil-to-lymphocyte ratio, serum LDH and C-reactive protein, as well as PD-L1 expression and MCPyV status, we could not identify a relevant association to therapy response. This observation is in-line with recently published study on 37 patients with MCC treated with CPI. Their multivariate analysis of clinical parameters including molecular subtype, age, prior radiotherapy, and PD-L1 expression did not show any predictive function (23).

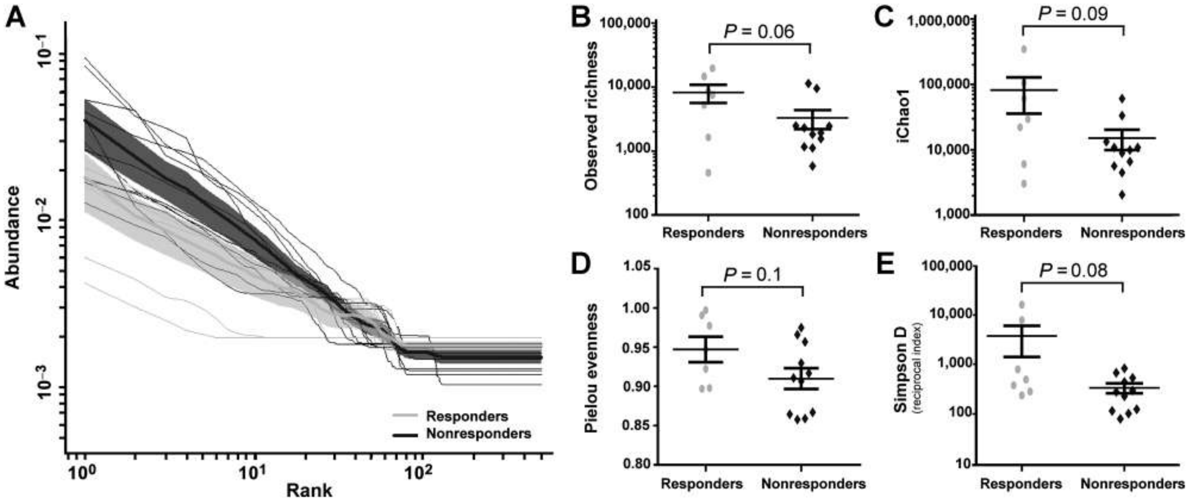


Figure 2.

High TCRB diversity among TILs predicts response to PD-1/PD-L1 blockade. **A,** Rank abundance distributions of TCRB clonotypes identified in TILs of responders ($n = 7$) and nonresponders ($n = 11$). Nonresponders are characterized by a larger expansion of T-cell clones. The rank-abundance distribution is normalized to the minimum TCRB richness and the number of individual TCRB clones. The averaged NRAD of each group is plotted as a bold line; shaded regions indicate 90% confidence intervals. **B,** Observed richness is a measure of different TCRB clones in each MCC tumor lesion. **C,** Estimated richness by iChao1 for rare clones. **D,** Pielou evenness reflects equal distribution of TCRB clones among specimens. **E,** Simpson diversity (Simpson D) represents the probability that two T cells taken at random from a specimen represent the same clone. *P* values were determined using the unpaired, two-tailed Student *t* test.

Lower T-cell clonality, but higher TCRB diversity of TILs is associated with response

Given the limited value of clinical, blood-based, and standard tissue-based biomarkers, we next performed a comprehensive, unbiased analysis to elucidate the impact of immunologic processes and immune cells characteristics on response to CPI. Given the need for sufficient amounts and quality of pretherapeutic tissue samples, these analyses were restricted to a subgroup of patients. The flow of the patients is detailed in Supplementary Fig. S1. To this end, 18 of the above-described patients with MCC were identified to meet these selection criteria to allow an in-depth immunologic and molecular work-up; this group consisted of seven responders and 11 nonresponders. Three of the 18 patients with MCC were stage III, not amenable to surgery or radiation and the other 15 were stage IV. Patients' characteristics are given in **Table 1**.

On the basis of the T-cell function and its attainment, the TCR repertoire usage is a mirror of immune response and differences in its usage have been suggested as a possible biomarker for the efficacy of immunomodulating therapies. Thus, we established the TCRB repertoire of TILs from MCC tumor samples obtained before and under anti–PD-1/PD-L1 therapy by high-throughput sequencing of the preamplified highly variable CDR3 sequences of the different TCR beta families. With their diverse repertoire of TCRs, T cells can be regarded as a generalized community, in which diversity reflects both richness and evenness. Using the Normalized Rank Abundance Distribution (NRAD; ref. 18), a descriptor for quantitative comparison of generalized communities, we compared the TCR repertoire usage with respect to response to CPI, revealing that TILs of responders had a more even clone size distribution, whereas in nonresponders the T-cell communities were dominated by a limited number of strongly expanded clones, corresponding to a high clonality (**Fig. 2A**). This observation is consistent with a higher richness for both strongly as well as weakly expanded T-cell clonotypes, the latter measured by the Chao index (**Fig. 2B** and **C**, respectively) and a higher evenness (**Fig. 2D**). Consequently, using the Simpson diversity reciprocal index to account for the clonal dominance hierarchy within each patient, TILs of responders is characterized by a higher diversity than those of nonresponders (**Fig. 2E**). Even though differences between both groups were not reckoned as significant by frequentist statistics, applying Bayes inference model supported the predictive value of the TCR repertoire richness and diversity for a favorable response to CPI treatment (Supplementary Fig. S2). The observed TCR repertoire richness showed the highest probability to be associated with therapy response.

T-cell attraction and activation genes are highly expressed in tumors of responders

Differences in TCR repertoire usage are likely to reflect functional differences of T cells. Analogous to other tissue types, T cells live through various stages of their differentiation, which are associated with variation of their function and proliferative capacity. Notably, these stages are discernable by gene expression. Thus, we performed NanoString-based gene expression analysis on 770 genes, characteristic for different immune cell types and their differential activation. A major advantage of these techniques, it is providing robust and reproducible results from FFPE tissues; however, sufficient integrity of RNA is still required to generate valid results. Thus, those pretherapeutic FFPE tissue samples that failed quality controls were excluded from analysis (Supplementary Fig. S1). In the remaining samples from $n = 6$ patients, the expression of immune-related genes clearly separated responders ($n = 3$) from nonresponders ($n = 3$; **Fig. 3A**; Supplementary Table S3). This discrimination was largely driven by approximately 100 genes involved in adaptive immunity, lymphocyte activation,

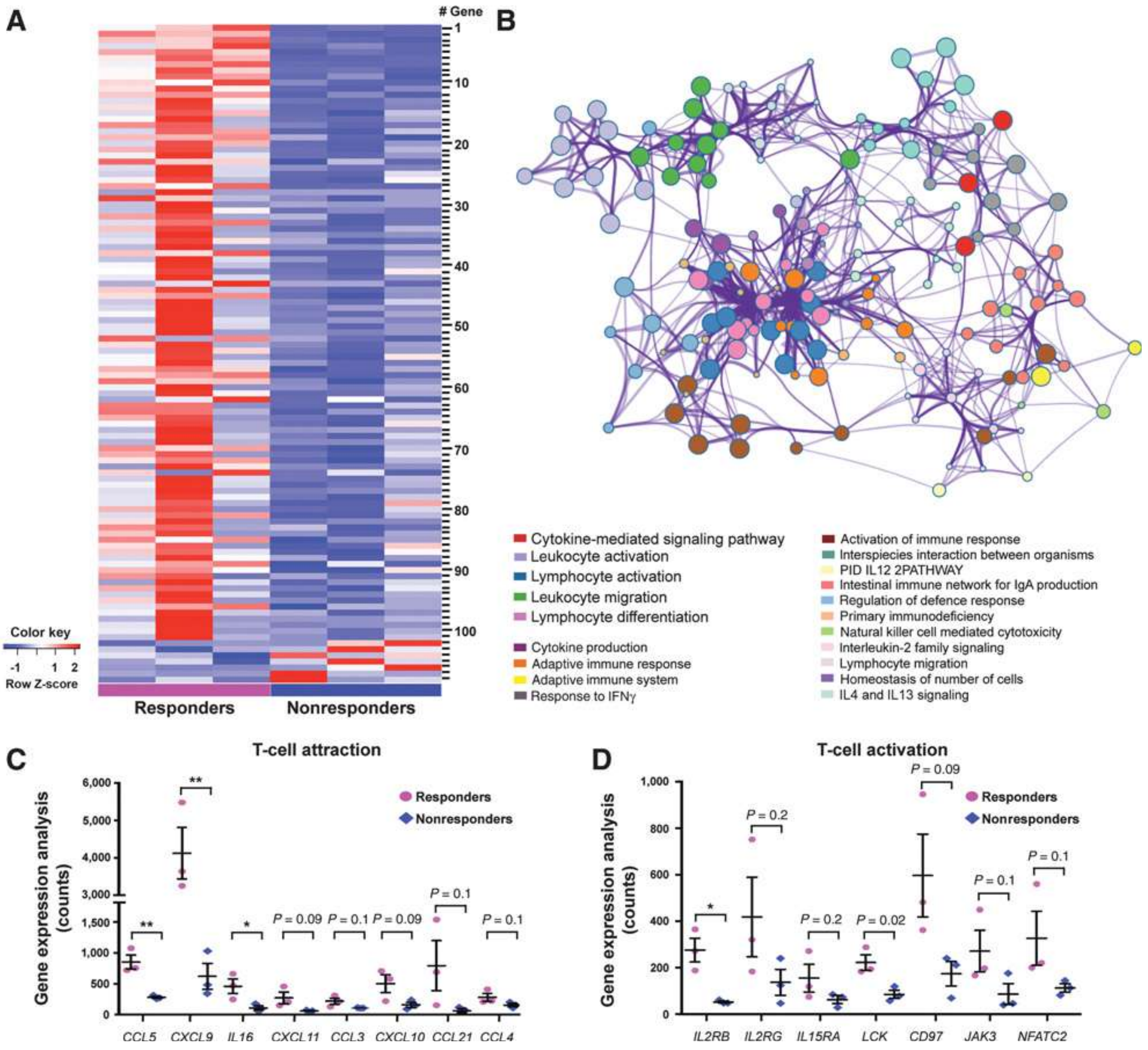


Figure 3.

High expression of genes related to T-cell attraction and activation predicts response to PD-1/PD-L1 blockade. NanoString gene expression analysis using the PanCancer Immune Profiling Panel at baseline from responders and nonresponders. **A,** Heatmap of differentially expressed genes; gene #1–101 are upregulated in responders and gene #102–108 are downregulated (details are given in Supplementary Table S3). Gene expression was normalized using the method of the geometric mean to the 32 most stable expressed genes. **B,** GO analysis of differentially expressed genes. The 20 top-score clusters with the lowest *P* values are depicted and each node represents an enriched term. The thickness of node connections represents term similarity, calculated by Cohen kappa coefficient. Expression level of selected genes related to T-cell attraction (**C**) and activation (**D**) for responders and nonresponders. *P* values were determined using the unpaired, two-tailed Student *t* test (*, $P < 0.05$; **, $P < 0.005$).

leukocyte migration, and cytokine signaling pathways as computed by GO analysis (**Fig. 3B**). In detail, genes related to T-cell attraction (e.g., *CCL5, CXCL9, IL16, CXCL11, CCL3, CXCL10, CCL21*, and *CCL4*; **Fig. 3C**) and T-cell activation (such as *IL2RB, IL2RG, IL15RA, LCK, CD97, JAK3*, and *NFATC2*; **Fig. 3D**) were highly expressed in tumor tissue from responders. Whereas the cell-cycle–related *CDK1* (#103) and the apoptosis regulation gene *BCL2* (#106) were strongly expressed in tumors of nonresponders (**Fig. 3A**). Notably, both genes have been linked to T-cell differentiation (24).

TILs of responders are dominated by central memory T cells, TILs of nonresponders by terminally differentiated T cells

The TCR repertoire usage and gene expression pattern in responders and nonresponders suggest disparate differentiation states of the respective predominant T-cell infiltrates. While memory T cells are characterized by their ability to exert a fast and sustained proliferative response to stimulation, exhausted and terminally differentiated T cells show an impaired proliferative capacity (25). Notably, for patients with melanoma it has been reported that predominance of memory T cells among TILs is associated with a favorable outcome of

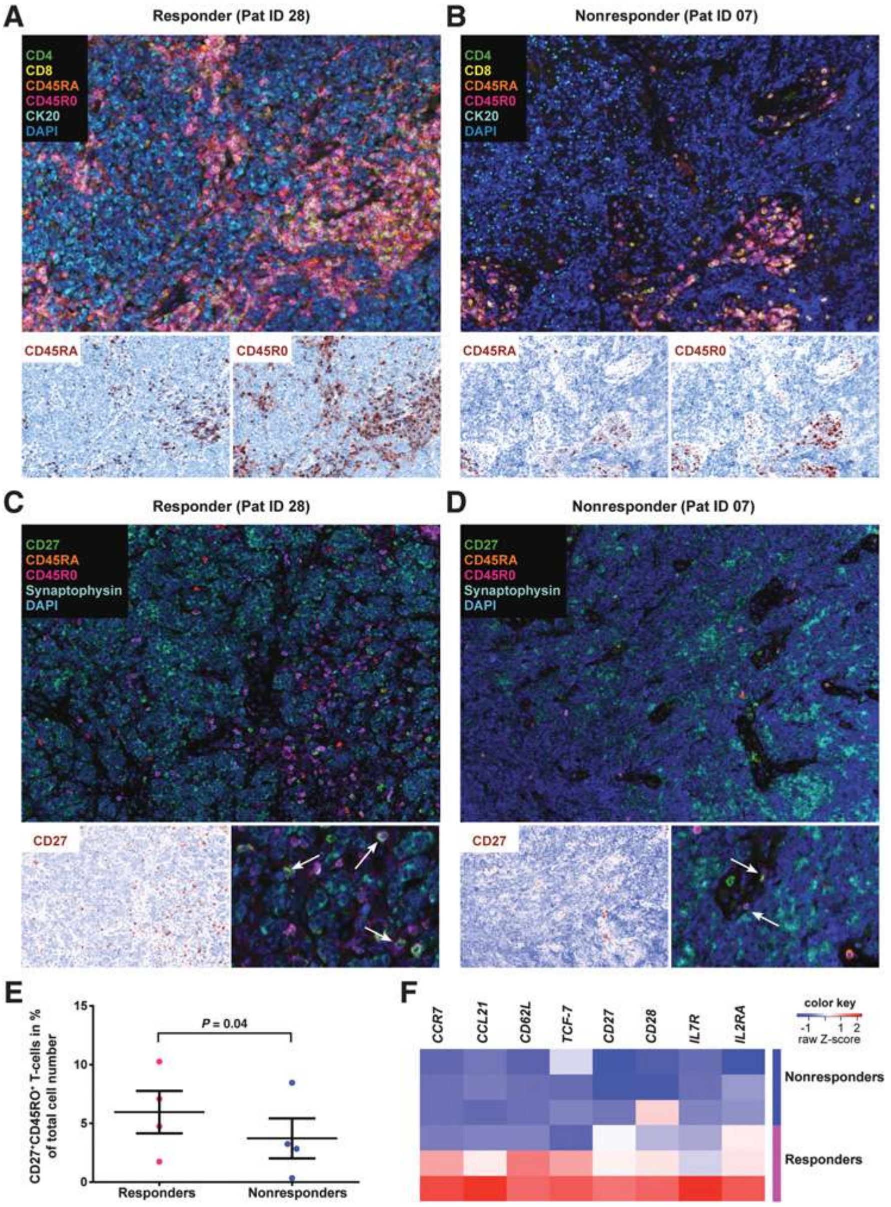


Figure 4.

Predominance of T_{CM} cells among TILs of responders. Multiplexed immunofluorescence staining of baseline tumor tissue obtained from responders (**A** and **C**) and nonresponders (**B** and **D**) using either antibodies against CD4 (green), CD8 (yellow), CD45RA (orange), CD45RO (magenta), and the MCC marker CK20 (cyan; **A** and **B**) or against CD27 (green), CD45RA (orange), CD45RO (magenta), and the MCC marker Synaptophysin (cyan; **C** and **D**); nuclei are stained with DAPI (blue). Depicted are merged images at 20 × magnification for all colors and single-channel images translated into pathology view for CD45RA and CD45RO or CD27, respectively, from one representative section. To visualize the colocalization of CD27 and CD45RO, an enlarged image is shown; white arrows highlight $CD27^{+}CD45RO^{+}$ T cells (**C** and **D**). **E,** Percentage of $CD27^{+}CD45RO^{+}$ T cells to total of nucleated cells for responders and nonresponders. **F,** Heatmap of expression of genes characteristic for T_{CM} cells in baseline tumor tissue.

CPI treatment (26–28). To test this hypothesis in our MCC patient cohort, we performed multiplexed immunofluorescence staining for memory ($CD45RO^{+}$) and effector ($CD45RA^{+}$) $CD4^{+}$ and $CD8^{+}$ T cells, revealing that in tumors of responding patients memory T cells had higher abundance and with respect to their spatial distribution they were located closer to the tumor cells, whereas in nonresponders such cells were rare and mostly present in the stromal compartments (**Fig. 4A** and **B**). A similar pattern was seen for effector T cells ($CD45RA^{+}$; **Fig. 4A** and **B**). To scrutinize the memory T cells in more detail, we included also the central memory T-cell marker CD27 in the staining panel ($n = 8$). The colocalization of CD27 with CD45RO demonstrated that T_{CM} cells are more frequent in responders (**Fig. 4C–E**; Supplementary Fig. S3; ref. 29). This notion was backed by gene expression signatures demonstrating higher abundance of genes such as *TCF-7, CCR7, CCL21, CD62L,* and *IL7R* in responders (**Fig. 4F**; refs. 29, 30).

Expression dynamics of genes related to lymphocyte activation, differentiation, migration, and presence of T_{CM} cells upon CPI treatment

Clustering of genes due to their functional relevance demonstrated that in responders, genes important for lymphocyte activation, differentiation, and migration, as well as cytokine-mediated signaling of T cells were not only more abundant at baseline, but also during therapy, indicating that this gene expression pattern was maintained or even boosted by CPI treatment (**Fig. 5A**). Furthermore, reflected by

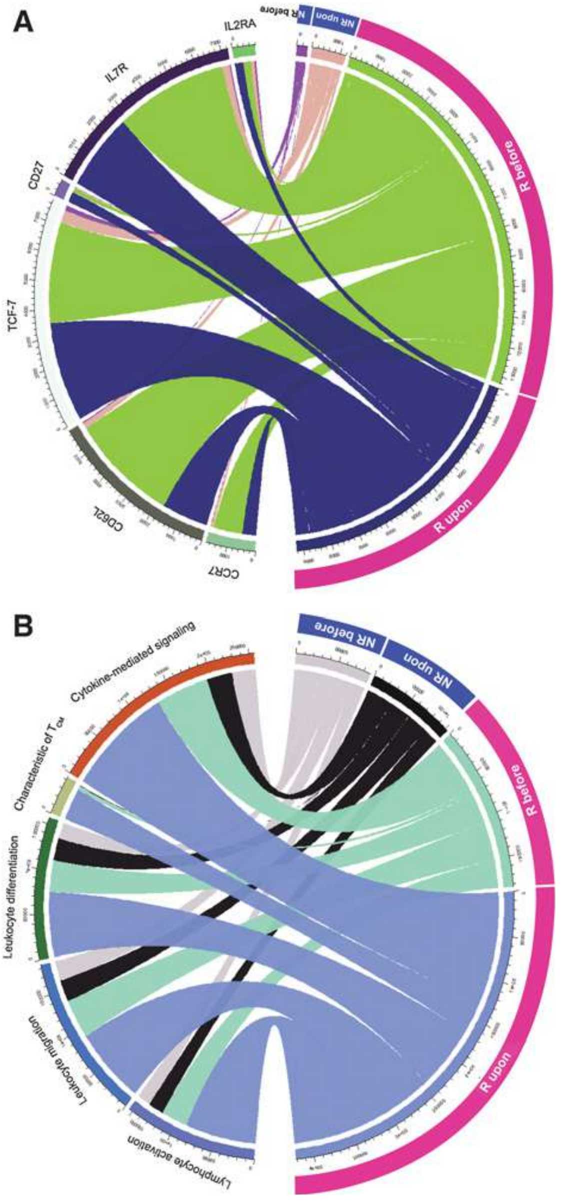


Figure 5.

Gene expression dynamics in MCC tumors upon PD-1/PD-L1 blockade. **A,** Circular plot displaying the expression dynamics of immune-related genes under CPI therapy. Lymphocyte activation, leukocyte migration, lymphocyte differentiation, and T$_{CM}$ cells and cytokine-mediated signaling genes are grouped together. The sum of expression of grouped genes in responders and nonresponders is depicted on the circular axis before (mint and grey) and upon (blue and black) CPI therapy, respectively. **B,** Circular plot displaying the expression dynamics of genes characteristic for T$_{CM}$ cells for an individual responder (Pat ID 28) and nonresponder (Pat ID 25) before and upon therapy. The expression of each gene is depicted on the circular axis before and upon anti–PD-1/PD-L1 therapy in green and blue (responder) and in purple and beige (nonresponder), respectively.

increased gene expression of *CCR7, CD62L, TCF-7, CD27, IL-7R*, and *IL-2RA* a higher presence of T$_{CM}$ cells was detected in tumor tissues, which also persisted upon therapy (**Fig. 5B**). In contrast, nonresponders showed no expression dynamics for the above-described genes upon treatment, that is, persistently low gene expression.

Dynamics of TCR repertoire usage upon PD-1/PD-L1 blockade

Next, we quantified the dynamics of TCR repertoire usage upon therapy (**Fig. 6A–F**). In accordance with the above-described results, the TCRB repertoire usage of TILs of a responder were distributed over a larger number of smaller clones than in a nonresponder. Notably, in the nonresponder the expansion of T-cell clones were up to one order of magnitude larger, indicating a substantial previous clonal expansion (**Fig. 6A** and **D**). On the other hand, upon PD-1/PD-L1 blockade, induced clonal T-cell expansions were more prominent among TILs of the responder (**Fig. 6A**, **B**, **D**, and **E**), suggesting a more pronounced proliferative capacity of T cells. Extending the comparative analyses of the TCRB repertoire from the dynamics of identical T-cell clones to newly emerging clones under therapy further confirmed that the number of newly emerging clones was higher among TILs of the responding patient (**Fig. 6C** and **F**).

An alternative way to construe the TCR repertoire usage is to cluster TCRs into convergence groups based on the CDR3 amino acid sequences that are predicted to bind the same or a similar MHC-restricted epitope using the recently published GLIPH algorithm (21). GLIPH predicts convergence groups calculating the probability that a cluster of similar TCRs has appeared by selection of a collectively recognized epitope/MHC complex. Representative examples for convergence group clustering by GLIPH for a responder and a nonresponder are depicted in **Fig. 6G** and **H**; for consistency, we chose the same two patients for which the multiplexed immunofluorescence staining results are presented. In the responder, three major TCRB clusters consisting of a multitude of different TCRs were present; this observation suggests that a large proportion of TILs characterized by a diverse TCR repertoire in responders are still recognizing a defined set of antigens. Notably, when we subjoined established MCPyV epitope-reactive TCR sequences, these were joined in two of these larger clusters (4). Of note, in both patients' tumors MCPyV DNA was detected (Supplementary Fig. S1). Upon treatment, there were no major changes in TCRB diversity and only some minor shifts in cluster formation (**Fig. 6G**). For the nonresponder, the limited TCR repertoire was characterized by almost complete absence of such convergence groups of TCR clonotypes. It should be noted that only TCRs within convergence groups, regardless of the individual expansion of a given TCR clonotype are depicted. After CPI treatment, the most obvious change was a further expansion of the two largest T-cell clones, which is represented by the increased size of the blue circles (**Fig. 6H**). Thus, in the patient with MCC with a favorable response to PD-1/PD-L1 blockade, we detected a substantial number of different T-cell clonotypes, which are recognizing a limited set of antigens. Results from the GLIPH analysis of further patients with MCC are depicted in Supplementary Fig. S4.

Discussion

The introduction of immunotherapy with CPI dramatically improved the hitherto poor prognosis of patients with advanced MCC. For either PD-1- or PD-L1–blocking antibodies, OR rates between 30% and 60% have been reported. Conversely, this means that almost half of the

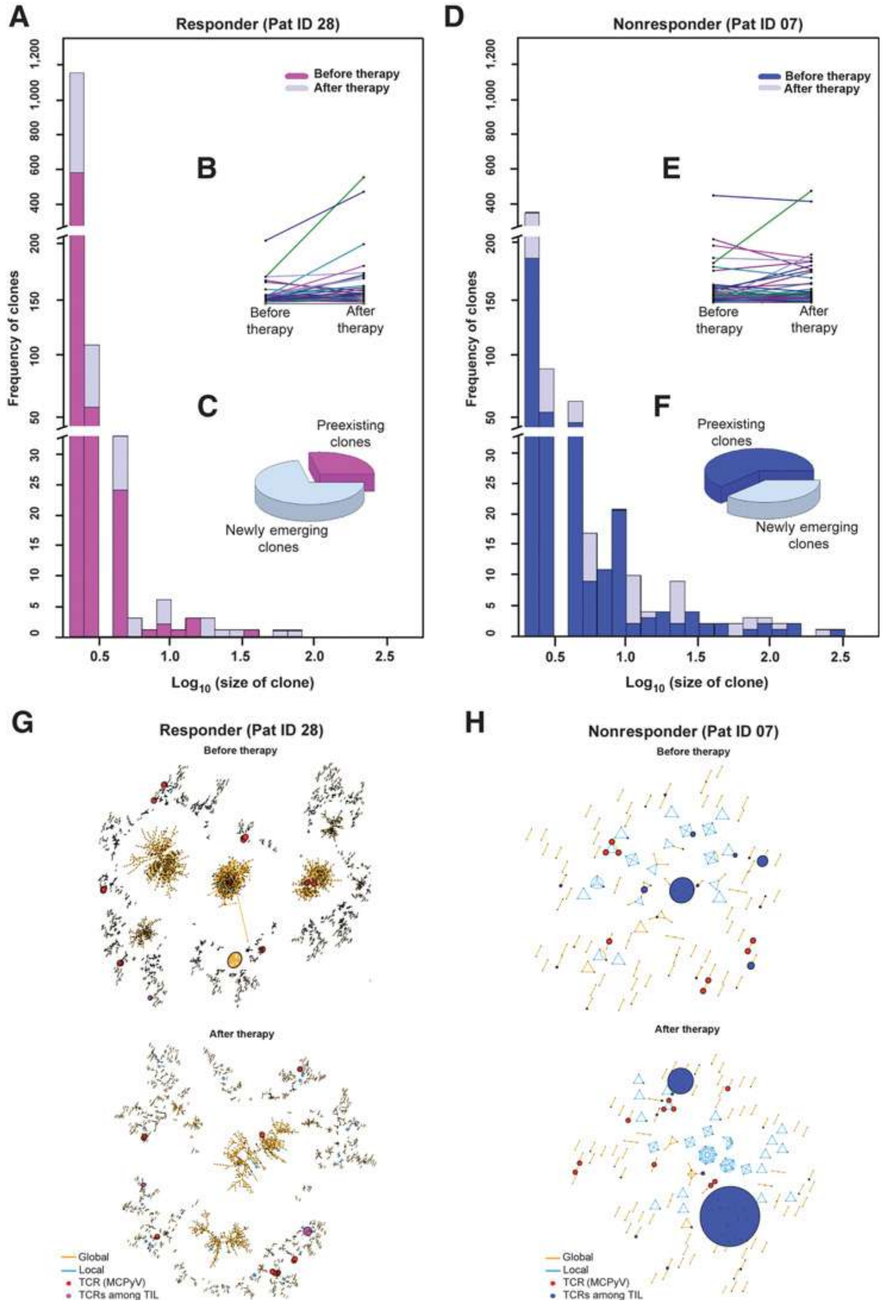


Figure 6.

Higher TCRB diversity recognizing a limited set of antigens is associated with response to PD-1/PD-L1 blockade. Frequency and size of TCRB clonotypes in TILs of a responding (**A**) and a nonresponding (**D**) patient before and upon treatment, frequencies are given as function of log-clone size. The TCRB frequency before therapy is depicted in a darker color, the frequency upon therapy in a lighter color, starting from the *x*-axis but being partly covered by the darker color. **B** and **E,** Frequencies in percent of the top 100 shared productive TCRB rearrangements before and under therapy. **C** and **F,** Pie charts for fractions of newly emerging TCRB clones as compared with preexisting clones. GLIPH analysis of TILs for TCRB clustering the CDR3 sequences into convergence groups assumed to react with the same peptide/MHC class I complex for a responder (**G**) and nonresponder (**H**). Global similarity of TCRs is represented by orange, local similarity by blue connecting lines. TCRB CDR3 sequences of previously established reactivity to known MCPyV epitopes are subjoined as red circles. Clone size is represented by the size of the magenta (responder) and blue (nonresponder) circles; only TCRBs within convergence groups are shown.

patients do not benefit from CPI and predictive biomarkers of therapy response are still lacking (6, 7, 31). In this study, we compiled clinical characteristics of 41 patients with MCC treated with CPI and tested currently presumed biomarkers, such as performance status, immunosuppression, previous therapies, serum LDH, neutrophil/lymphocyte ratio, PD-L1 expression, and MCPyV status. We chose to apply a Bayesian logistic regression model instead of classical least squared regression due to the given study character. Specifically, the study comprises a limited sample size and a large number of variables to be analyzed for potential effects on treatment response. In classical statistics these features lead to overfitting (14). Among the analyzed parameters, only performance status and immune suppression correlated with response to therapy. Thus, we next performed a comprehensive immunologic work-up of MCC tumor tissues taken at therapy baseline and additionally in a subgroup of patients after initiation of therapy. This work-up comprised high-throughput sequencing-based TCR clonotype mapping, multiplexed immunofluorescence staining, and immune gene mRNA expression. Thereby, we established that a prevalence of T_{CM} cells expressing a highly diverse TCRB repertoire among TILs is associated with a favorable therapy response. Notably, even though these parameters were established from only a subgroup, Bayes inference revealed that their predictive value is comparable with good performance status and absence of immune suppression of the complete group (Supplementary Fig. S2).

Because clonal expansion is one fundamental event in effector T-cell development, the TCR repertoire reflects both the previous history and future prospects of adaptive cellular immune responses (32). The TCRB repertoire of responders was characterized by higher richness, an indicator accounting not only for the number of individual TCRBs, but also for their heterogeneity. Greater richness was associated with a higher evenness of TCRB clonotype, thus resulting in a high T-cell clonotype diversity. The nature of this T-cell diversity was further analyzed by clustering the respective TCRB CDR3 regions by similarity. GLIPH analysis revealed that in responders the highly diverse TCR repertoire was readily grouped in a limited number of convergence groups. Notably, when artificially joining in established TCR CDR3 sequences of T cells specifically reacting with MCPyV epitopes (4), these were also clustered within these convergence groups. In TILs of nonresponding patients, despite larger expansion of individual clones, such convergence groups were virtually absent. The positive predictive value of a diverse TCR repertoire for response to CPI therapy was recently also observed in melanoma (33). Consistent with the predictive value of a high TCR-repertoire evenness, our NRAD analysis demonstrated that TILs of nonresponders were dominated by a few, largely expanded T-cell clones. Thus, our findings indicate that in responding patients, immune-inhibiting conditions prevented clonal expansion of T cells, which could be abrogated by anti–PD-1/PD-L1 therapy, whereas in nonresponders' reactive T cells were previously expanded reaching their proliferative capacity, thus a state not amendable by immune checkpoint blockade. To test this, we additionally addressed dynamic changes of the TCR repertoire usage upon CPI therapy revealing that both size and number of T-cell clones increased in responders, whereas in nonresponders the already large T-cell clones at baseline did not undergo the same relative expansion and the number of newly emerging clones was substantially smaller. Moreover, the TCR-repertoire diversity in nonresponders remained low, suggesting an irreversible T-cell dysfunction, for example, by terminal differentiation. To this end, the age-related involution of the thymus is associated with a shift of the T-cell pool from naïve to effector memory T cells (34, 35). As a result, the naïve T-cell repertoire is increasingly curtailed, which is evidenced by a loss of TCR-repertoire diversity (36). The notion appears to be particularly important for MCC, characterized by an elderly patient population (1). These observations stress the importance of the TCR-repertoire diversity as a pivotal feature of a functional immune system. In-line with this concept, functional characterization of the immune microenvironment in MCC before initiation of CPI treatment by gene expression analyses demonstrated a very low or no expression of genes related to T-cell attraction or activation in nonresponders; on the other hand, these genes were highly expressed in tumors of responders. Similarly, recent studies of TILs in melanoma demonstrated a correlation of TCR clonality and the fraction of dysfunctional T cells (37).

Deconvolution of the molecular immune phenotype derived from gene expression data revealed higher numbers of T_{CM} cells in responding patients. Immunofluorescence detection of high numbers of $CD27^+$ $CD45RO^+$ T cells among TILs of responders confirmed this notion. The amount of T_{CM} cells remained stable upon CPI treatment, which was also observed in a preclinical therapy model of colon cancer (38). This maintenance of T_{CM} cells may be explained by the strong expression of the transcription factor *TCF7*, which is crucial for differentiation, self-renewal, and persistence of memory $CD8^+$ T cells. Indeed, *TCF7* expression remained high during treatment. Similarly, studies in melanoma demonstrated that *TCF7* is linked to an effective $CD8^+$ T-cell response upon immunotherapy and elevated frequencies of $TCF7^+CD8^+$ T cell are predictive for therapy response (38, 39). In contrast, TILs of nonresponders were characterized by a high prevalence of terminally differentiated, nonfunctional T cells, characterized by expression of *BCL-2*. Terminally differentiated T cells have to be distinguished from exhausted T cells. T-cell exhaustion is a consequence of continuous stimulation causing a gradual loss of effector capabilities and expression of inhibitory receptors. Exhausted T cells, however, can be readily reactivated by therapeutic interventions such as inhibition of PD-1/PD-L1 pathway (40). Terminally differentiated T cells, in contrast, are induced by overstimulation causing excessive proliferation resulting in critically shortened telomeres (41–43). Thus, terminally differentiated T cells have reached the final phase of their activation cycle, and can no longer be reactivated even by immune checkpoint blockade or epigenetic modifiers (29).

In conclusion, we identified immunologic and molecular characteristics measurable in tumor tissue at treatment baseline associated with clinical response to anti–PD-1/PD-L1 therapy of advanced MCC. In patients with a favorable response, TILs had a rich and diverse TCR repertoire as well as a phenotype of T_{CM} cells, whereas a predominance of largely expanded, terminally differentiated T cells was associated with an impaired response.

Disclosure of Potential Conflicts of Interest

S. Ugurel is an employee/paid consultant for Bristol-Myers Squibb, Merck, and Merck Sharp and Dohme. P. Terheyden reports receiving other remuneration from Sanofi, Pfizer, Roche, Bristol-Myers Squibb, Novartis, Roche, and Biofrontera. J.C. Hassel is an employee/paid consultant for Merck Sharp and Dohme, Pierre Fabre, Sanofi, and Sun Pharma, and is an advisory board member/unpaid consultant for Bristol-Myers Squibb and Novartis. J.C. Becker is an employee/paid consultant for Sanofi, Merck Serono, ReProTher, 4SC, Amgen, eTheRNA, and Pfizer, reports receiving commercial research grants from Alcedis, IQVIA, Merck Serono, and Amgen, speakers bureau honoraria from Amgen, Merck Serono, Sanofi, Pfizer, and Recordati, and other remuneration from Incyte. No potential conflicts of interest were disclosed by the other authors.

Authors' Contributions

Conception and design: I. Spassova, S. Ugurel, J.C. Becker
Development of methodology: I. Spassova, L. Kubat, L. Peiffer, J.C. Becker
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): S. Ugurel, P. Terheyden, A. Sucker, J.C. Hassel, C. Ritter, R. Kumar, D. Schadendorf, J.C. Becker
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): I. Spassova, S. Ugurel, C. Ritter, D. Habermann, F. Farahpour, M. Saeedghalati, L. Peiffer, R. Kumar, D. Schrama, D. Hoffmann, J.C. Becker
Writing, review, and/or revision of the manuscript: I. Spassova, S. Ugurel, P. Terheyden, J.C. Hassel, C. Ritter, L. Peiffer, R. Kumar, D. Schrama, D. Hoffmann, D. Schadendorf, J.C. Becker
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): I. Spassova, S. Ugurel, A. Sucker, J.C. Becker
Study supervision: R. Kumar, J.C. Becker

Acknowledgments

This work was funded by the DKTK site budget OE 0460 EDO3. We thank the patients who participated in this study, their families, and the staff members at the clinical centers who cared for them.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received July 9, 2019; revised November 6, 2019; accepted January 8, 2020; published first January 13, 2020.

References

1. Becker JC, Stang A, DeCaprio JA, Cerroni L, Lebbe C, Veness M, et al. Merkel cell carcinoma. Nat Rev Dis Primers 2017;3:17077.
2. Paulson KG, Iyer JG, Blom A, Warton EM, Sokil M, Yelistratova L, et al. Systemic immune suppression predicts diminished Merkel cell carcinoma-specific survival independent of stage. J Invest Dermatol 2013;133:642–6.
3. Feng H, Shuda M, Chang Y, Moore PS. Clonal integration of a polyomavirus in human Merkel cell carcinoma. Science 2008;319:1096–100.
4. Lyngaa R, Pedersen NW, Schrama D, Thrue CA, Ibrani D, Met O, et al. T-cell responses to oncogenic Merkel cell polyomavirus proteins distinguish patients with Merkel cell carcinoma from healthy donors. Clin Cancer Res 2014;20:1768–78.
5. Harms PW, Vats P, Verhaegen ME, Robinson DR, Wu YM, Dhanasekaran SM, et al. The distinctive mutational spectra of polyomavirus-negative Merkel cell carcinoma. Cancer Res 2015;75:3720–7.
6. Nghiem PT, Bhatia S, Lipson EJ, Kudchadkar RR, Miller NJ, Annamalai L, et al. PD-1 blockade with pembrolizumab in advanced Merkel-cell carcinoma. N Engl J Med 2016;374:2542–52.
7. D'Angelo SP, Russell J, Lebbe C, Chmielowski B, Gambichler T, Grob JJ, et al. Efficacy and safety of first-line avelumab treatment in patients with stage IV metastatic Merkel cell carcinoma: a preplanned interim analysis of a clinical trial. JAMA Oncol 2018;4:e180077.
8. Heidelberger V, Goldwasser F, Kramkimel N, Jouinot A, Franck N, Arrondeau J, et al. Clinical parameters associated with anti-programmed death-1 (PD-1) inhibitors-induced tumor response in melanoma patients. Invest New Drugs 2017;35:842–7.
9. Maleki Vareki S, Garrigos C, Duran I. Biomarkers of response to PD-1/PD-L1 inhibition. Crit Rev Oncol Hematol 2017;116:116–24.
10. Kaufman HL, Russell J, Hamid O, Bhatia S, Terheyden P, D'Angelo SP, et al. Avelumab in patients with chemotherapy-refractory metastatic Merkel cell carcinoma: a multicentre, single-group, open-label, phase 2 trial. Lancet Oncol 2016;10:1374–85.
11. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer 2009;45:228–47.
12. Becker JC, Houben R, Ugurel S, Trefzer U, Pfohler C, Schrama D. MC polyomavirus is frequently present in Merkel cell carcinoma of European patients. J Invest Dermatol 2009;129:248–50.
13. Wolchok JD, Kluger H, Callahan MK, Postow MA, Rizvi NA, Lesokhin AM, et al. Safety and clinical activity of combined PD-1 (nivolumab) and CTLA-4 (ipilimumab) blockade in advanced melanoma patients. N Engl J Med 2013;369: 122–33.
14. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian data analysis. United Kingdom: Chapman and Hall/CRC; 2013.
15. Bürkner PC. brms: an R package for Bayesian multilevel models using stan. J Stat Softw 2017;80:1–28.
16. Bürkner PC. Advanced Bayesian multilevel modeling with the R package brms. The R Journal 2018;10:395–411.
17. van Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. J Stat Softw 2011;45:1–67.
18. Saeedghalati M, Farahpour F, Budeus B, Lange A, Westendorf AM, Seifert M, et al. Quantitative comparison of abundance structures of generalized communities: from B-cell receptor repertoires to microbiomes. PLoS Comput Biol 2017; 13:e1005362.
19. Mansfield AS, Ren H, Sutor S, Sarangi V, Nair A, Davila J, et al. Contraction of T cell richness in lung cancer brain metastases. Sci Rep 2018;8:2171.
20. Chiu CH, Wang YT, Walther BA, Chao A. An improved nonparametric lower bound of species richness via a modified good-turing frequency formula. Biometrics 2014;70:671–82.
21. Glanville J, Huang H, Nau A, Hatton O, Wagar LE, Rubelt F, et al. Identifying specificity groups in the T cell receptor repertoire. Nature 2017;547:94–8.
22. Tripathi S, Pohl MO, Zhou Y, Rodriguez-Frandsen A, Wang G, Stein DA, et al. Meta- and orthogonal integration of influenza "OMICs" data defines a role for UBR4 in virus budding. Cell Host Microbe 2015;18:723–35.
23. Knepper TC, Montesion M, Russell JS, Sokol ES, Frampton GM, Miller VA, et al. The genomic landscape of Merkel cell carcinoma and clinicogenomic biomarkers of response to immune checkpoint inhibitor therapy. Clin Cancer Res 2019:5961–71.
24. Wells AD, Morawski PA. New roles for cyclin-dependent kinases in T cell biology: linking cell division and differentiation. Nat Rev Immunol 2014; 14:261–70.
25. Jin G, Xue G, Wang R-S, Wu L-Y, Miller L, Lu Y, et al. Single-cell modeling of CD8+ T cell exhaustion predicts response to cancer immunotherapy. bioRxiv 2018.
26. Takeuchi Y, Tanemura A, Tada Y, Katayama I, Kumanogoh A, Nishikawa H. Clinical response to PD-1 blockade correlates with a sub-fraction of peripheral central memory CD4+ T cells in patients with malignant melanoma. Int Immunol 2018;30:13–22.
27. Krieg C, Nowicka M, Guglietta S, Schindler S, Hartmann FJ, Weber LM, et al. High-dimensional single-cell analysis predicts response to anti-PD-1 immunotherapy. Nat Med 2018;24:144–53.
28. Edwards J, Wilmott JS, Madore J, Gide TN, Quek C, Tasker A, et al. CD103(+) tumor-resident CD8(+) T cells are associated with improved survival in immunotherapy-naive melanoma patients and expand significantly during anti-PD-1 treatment. Clin Cancer Res 2018;24:3036–45.
29. Mahnke YD, Brodie TM, Sallusto F, Roederer M, Lugli E. The who's who of T-cell differentiation: human memory T-cell subsets. Eur J Immunol 2013;43:2797–809.
30. Danilo M, Chennupati V, Silva JG, Siegert S, Held W. Suppression of Tcf1 by inflammatory cytokines facilitates effector CD8 T cell differentiation. Cell Rep 2018;22:2107–17.
31. Giraldo NA, Nguyen P, Engle EL, Kaunitz GJ, Cottrell TR, Berry S, et al. Multidimensional, quantitative assessment of PD-1/PD-L1 expression in patients with Merkel cell carcinoma and association with response to pembrolizumab. J Immunother Cancer 2018;6:99.
32. Schrama D, Ritter C, Becker JC. T cell receptor repertoire usage in cancer as a surrogate marker for immune responses. Semin Immunopathol 2017;39:255–68.
33. Postow MA, Manuel M, Wong P, Yuan J, Dong Z, Liu C, et al. Peripheral T cell receptor diversity is associated with clinical outcomes following ipilimumab treatment in metastatic melanoma. J Immunother Cancer 2015;3:23.
34. Alpert A, Pickman Y, Leipold M, Rosenberg-Hasson Y, Ji X, Gaujoux R, et al. A clinically meaningful metric of immune age derived from high-dimensional longitudinal monitoring. Nat Med 2019;25:487–95.
35. Surh CD, Sprent J. Homeostasis of naive and memory T cells. Immunity 2008;29: 848–62.
36. Mekker A, Tchang VS, Haeberli L, Oxenius A, Trkola A, Karrer U. Immune senescence: relative contributions of age and cytomegalovirus infection. PLoS Pathog 2012;8:e1002850.
37. Li H, van der Leun AM, Yofe I, Lubling Y, Gelbard-Solodkin D, van Akkooi ACJ, et al. Dysfunctional CD8 T cells form a proliferative, dynamically regulated compartment within human melanoma. Cell 2019;176:775–89.
38. Kurtulus S, Madi A, Escobar G, Klapholz M, Nyman J, Christian E, et al. Checkpoint blockade immunotherapy induces dynamic changes in PD-1(-)CD8 (+) tumor-infiltrating T cells. Immunity 2019;50:181–94.
39. Sade-Feldman M, Yizhak K, Bjorgaard SL, Ray JP, de Boer CG, Jenkins RW, et al. Defining T cell states associated with response to checkpoint immunotherapy in melanoma. Cell 2018;175:998–1013.
40. Kamphorst AO, Wieland A, Nasti T, Yang S, Zhang R, Barber DL, et al. Rescue of exhausted CD8 T cells by PD-1-targeted therapies is CD28-dependent. Science 2017;355:1423–7.
41. Zarour HM. Reversing T-cell dysfunction and exhaustion in cancer. Clin Cancer Res 2016;22:1856–64.
42. Paley MA, Kroy DC, Odorizzi PM, Johnnidis JB, Dolfi DV, Barnett BE, et al. Progenitor and terminal subsets of CD8+ T cells cooperate to contain chronic viral infection. Science 2012;338:1220–5.
43. Odorizzi PM, Pauken KE, Paley MA, Sharpe A, Wherry EJ. Genetic absence of PD-1 promotes accumulation of terminally differentiated exhausted CD8+ T cells. J Exp Med 2015;212:1125–37.

Chapter 4

Discussion

The contributed articles show that Bayesian statistics can be useful for a wide range of applications: In `HAMdetector`, Bayesian statistics is used mainly as a tool to integrate information from different sources. Phylogeny, epitope prediction and HLA alleles all add different pieces of information that can be combined in a larger model that accounts for uncertainty in the inputs. The phylogenetic tree is used to relax the assumption of independent samples, epitope prediction provides information on the a-priori probability of a given replacement being HLA associated, and the observed counts of replacements and HLA alleles provide the ”hard data” to update the likelihood.

In the MCC articles, Bayesian statistics is used as a flexible tool to adapt a statistical model to the observed data. In Spassova et al. (2020), the limited data led us to fit a model that ignores the ordinal character of the categories, and we therefore fit a logistic regression model that bins the categories ”progressive response” and ”partial response” into a ”response” group, and the categories ”stable disease” and ”progressive disease” into a ”non-response” group. While binning does go against the general principle of including as much data as possible into a statistical model, in this case it can be seen as imposing the assumption that the categories in each bin behave similarly as a function of the predictors. The resulting posterior intervals for the regression coefficients are necessarily broad, but the data did suggest that some risk factors are more strongly linked to therapy non-response than others. As more data was collected (Spassova et al., 2022), it became possible to fit a larger model that includes the ordinal character of the data. Such models are difficult to fit in a non-Bayesian setting, as some regularization helps to reduce overestimating possible effects, and it is also relatively easy to account for missing data in some of the inputs.

To summarize, the contributed articles highlight Bayesian statistics in a wide range of applications: Habermann et al. (2022) focuses on the aspect of integrating information from different sources, whereas Spassova et al. (2020) shows that it is possible to get reasonable inferences when data are sparse and noisy. This is achieved by applying methods that are skeptical by default (e.g. by using a regularizing prior), and account for missing data in such a way that uncertainty in the inputs translates into uncertainty of the inferences. Additionally, Spassova et al. (2022) also highlights some model visualization techniques: Regression coefficients for models with non-linearities like logistic regression are difficult to interpret, as they make statements about predictors on the log-odds scale. Average predictive comparisons instead make statements about the expected change in the outcome with a unit difference in one of the inputs (Gelman and Pardoe, 2007). This can help to present results of statistical models in a way that is more easily interpretable. While this idea is not limited to Bayesian models per se, describing model parameters with probability distributions has the advantage that we can also express average predictive comparisons in terms of distributions.

Bibliography

Rupert Abele and Robert Tampé. The ABCs of immunology: Structure and function of TAP, the transporter associated with antigen processing. *Physiology*, 19(4):216–224, aug 2004. doi: 10.1152/physiol.00002.2004.

Erin J. Adams, Siyi Gu, and Adrienne M. Luoma. Human gamma delta T cells: Evolution and ligand recognition. *Cellular Immunology*, 296(1):31–40, jul 2015. doi: 10.1016/j.cellimm.2015.04.008.

Julian Adams. The proteasome: structure, function, and role in the cell. *Cancer Treatment Reviews*, 29:3–9, may 2003. doi: 10.1016/s0305-7372(03)00081-1.

Aimé Cézaire Adiko, Joel Babdor, Enric Gutiérrez-Martínez, Pierre Guermonprez, and Loredana Saveanu. Intracellular transport routes for MHC I and their relevance for antigen cross-presentation. *Frontiers in Immunology*, 6, jul 2015. doi: 10.3389/fimmu.2015.00335.

D. G Altman and J M. Bland. Statistics notes: Absence of evidence is not evidence of absence. *BMJ*, 311(7003):485–485, aug 1995. doi: 10.1136/bmj.311.7003.485.

M. J. Androlewicz, K. S. Anderson, and P. Cresswell. Evidence that transporters associated with antigen processing translocate a major histocompatibility complex class I-binding peptide into the endoplasmic reticulum in an ATP-dependent manner. *Proceedings of the National Academy of Sciences*, 90(19):9130–9134, oct 1993. doi: 10.1073/pnas.90.19.9130.

Matthew J. Androlewicz and Peter Cresswell. Human transporters associated with antigen processing possess a promiscuous peptide-binding site. *Immunity*, 1(1):7–14, apr 1994. doi: 10.1016/1074-7613(94)90004-3.

E. K. Barber, J. D. Dasgupta, S. F. Schlossman, J. M. Trevillyan, and C. E. Rudd. The CD4 and CD8 antigens are coupled to a protein-tyrosine kinase (p56lck) that phosphorylates the CD3 complex. *Proceedings of the National Academy of Sciences*, 86 (9):3277–3281, may 1989. doi: 10.1073/pnas.86.9.3277.

Jared A.M. Bard, Ellen A. Goodall, Eric R. Greene, Erik Jonsson, Ken C. Dong, and Andreas Martin. Structure and function of the 26S proteasome. *Annual Review of*

Biochemistry, 87(1):697–724, jun 2018. doi: 10.1146/annurev-biochem-062917-011931. Proteasome review.

Charles K Barlowe and Elizabeth A Miller. Secretory protein biogenesis and traffic in the early secretory pathway. *Genetics*, 193(2):383–410, feb 2013. doi: 10.1534/genetics.112.142810.

Craig H Bassing, Wojciech Swat, and Frederick W Alt. The mechanism and regulation of chromosomal v(d)j recombination. *Cell*, 109(2):S45–S55, apr 2002. doi: 10.1016/s0092-8674(02)00675-x.

Susann Beetz, Daniela Wesch, Lothar Marischen, Stefan Welte, Hans-Heinrich Oberg, and Dieter Kabelitz. Innate immune functions of human $\gamma\delta$ T cells. *Immunobiology*, 213(3-4):173–182, may 2008. doi: 10.1016/j.imbio.2007.10.006.

Michael Betancourt. Probability theory (for scientists and engineers), 2018. URL `https://betanalpha.github.io/assets/case_studies/probability_theory.html#1_setting_a_foundation`. Accessed on 09.03.2022.

Michael Betancourt. Principled bayesian workflow, 2020. URL `https://betanalpha.github.io/assets/case_studies/principled_bayesian_workflow`. Accessed on 26.03.2022.

Michael Betancourt. Prior modeling, 2021. URL `https://betanalpha.github.io/assets/case_studies/prior_modeling.html`. Accessed on 12.03.2022.

Purnima Bhat, Graham Leggatt, Klaus I. Matthaei, and Ian H. Frazer. The kinematics of cytotoxic lymphocytes influence their ability to kill target cells. *PLoS ONE*, 9(5): e95248, may 2014. doi: 10.1371/journal.pone.0095248.

Purnima Bhat, Graham Leggatt, Nigel Waterhouse, and Ian H Frazer. Interferon-γ derived from cytotoxic lymphocytes directly enhances their motility and cytotoxicity. *Cell Death & Disease*, 8(6):e2836–e2836, jun 2017. doi: 10.1038/cddis.2017.67.

P. J. Bjorkman, M. A. Saper, B. Samraoui, W. S. Bennett, J. L. Strominger, and D. C. Wiley. Structure of the human class I histocompatibility antigen, HLA-a2. *Nature*, 329 (6139):506–512, oct 1987. doi: 10.1038/329506a0.

Andreas Blees, Katrin Reichel, Simon Trowitzsch, Olivier Fisette, Christoph Bock, Rupert Abele, Gerhard Hummer, Lars V. Schäfer, and Robert Tampé. Assembly of the MHC I peptide-loading complex determined by a conserved ionic lock-switch. *Scientific Reports*, 5(1), nov 2015. doi: 10.1038/srep17341.

Ljudmila Borissenko and Michael Groll. 20S proteasome and its inhibitors: crystallographic knowledge for drug development. *Chemical Reviews*, 107(3):687–717, feb 2007. doi: 10.1021/cr0502504.

M Bouvier and D. Wiley. Importance of peptide amino and carboxyl termini to the stability of MHC class I molecules. *Science*, 265(5170):398–402, jul 1994. doi: 10.1126/science.8023162.

Marlene Bouvier. Accessory proteins and the assembly of human class I MHC molecules: a molecular and structural perspective. *Molecular Immunology*, 39(12):697–706, jan 2003. doi: 10.1016/s0161-5890(02)00261-4.

Michael A. Brehm, Keith A. Daniels, and Raymond M. Welsh. Rapid production of TNF-α following TCR engagement of naive CD8 T Cells. *The Journal of Immunology*, 175(8):5043–5049, oct 2005. doi: 10.4049/jimmunol.175.8.5043.

P. Bretscher and M. Cohn. A theory of self-nonself discrimination: Paralysis and induction involve the recognition of one and two determinants on an antigen, respectively. *Science*, 169(3950):1042–1049, sep 1970. doi: 10.1126/science.169.3950.1042.

Lori Buetow and Danny T. Huang. Structural insights into the catalysis and regulation of E3 ubiquitin ligases. *Nature Reviews Molecular Cell Biology*, 17(10):626–642, aug 2016. doi: 10.1038/nrm.2016.91.

Trevor D. Burt. Fetal regulatory T cells and peripheral immune Tolerance In utero: Implications for development and disease. *American Journal of Reproductive Immunology*, 69(4):346–358, feb 2013. doi: 10.1111/aji.12083.

Rasmus Bååth. Diy kruschke style diagrams, 2016. URL `https://www.sumsar.net/blog/2013/10/diy-kruschke-style-diagrams/`. Accessed on 13.03.2022.

Sidney B. Cambridge, Florian Gnad, Chuong Nguyen, Justo Lorenzo Bermejo, Marcus Krüger, and Matthias Mann. Systems-wide proteomic analysis in mammalian cells reveals conserved, functional protein turnover. *Journal of Proteome Research*, 10(12): 5275–5284, dec 2011. doi: 10.1021/pr101183k.

Alexander Chernev, Ulf Böckenholt, and Joseph Goodman. Choice overload: A conceptual review and meta-analysis. *Journal of Consumer Psychology*, 25(2):333–358, apr 2015. doi: 10.1016/j.jcps.2014.08.002.

Sunglim Cho, Jeongmin Ryoo, Youngsoo Jun, and Kwangseog Ahn. Receptor-mediated ER export of human MHC class I molecules is regulated by the c-terminal single amino acid. *Traffic*, 12(1):42–55, nov 2010. doi: 10.1111/j.1600-0854.2010.01132.x.

R. T. Cox. Probability, frequency and reasonable expectation. *American Journal of Physics*, 14(1):1–13, jan 1946. doi: 10.1119/1.1990764.

Weiguo Cui and Susan M. Kaech. Generation of effector CD8+ T cells and their conversion to memory T cells. *Immunological Reviews*, 236(1):151–166, jun 2010. doi: 10.1111/j.1600-065x.2010.00926.x.

Julie M. Curtsinger, Christopher M. Johnson, and Matthew F. Mescher. CD8 T cell clonal expansion and development of effector function require prolonged exposure to antigen, costimulation, and signal 3 cytokine. *The Journal of Immunology*, 171(10):5165–5171, nov 2003. doi: 10.4049/jimmunol.171.10.5165.

Bruno de Finetti. *Theory of Probability*. Wiley Series in Probability and Statistics. John Wiley & Sons, Nashville, TN, March 1974.

E Degen and D B Williams. Participation of a novel 88-kD protein in the biogenesis of murine class I histocompatibility molecules. *Journal of Cell Biology*, 112(6):1099–1115, mar 1991. doi: 10.1083/jcb.112.6.1099.

Ineke den Braber, Tendai Mugwagwa, Nienke Vrisekoop, Liset Westera, Ramona Mögling, Anne Bregje de Boer, Neeltje Willems, Elise H.R. Schrijver, Gerrit Spierenburg, Koos Gaiser, Erik Mul, Sigrid A. Otto, An F.C. Ruiter, Mariette T. Ackermans, Frank Miedema, José A.M. Borghans, Rob J. de Boer, and Kiki Tesselaar. Maintenance of peripheral naive T cells is sustained by thymus output in mice but not humans. *Immunity*, 36(2):288–297, feb 2012. doi: 10.1016/j.immuni.2012.02.006.

Alex Deng, Jiannan Lu, and Shouyuan Chen. Continuous monitoring of a/b tests without pain: Optional stopping in bayesian testing, 2016.

Sarah Enouz, Lucie Carrié, Doron Merkler, Michael J. Bevan, and Dietmar Zehn. Autoreactive T cells bypass negative selection and respond to self-antigen stimulation during infection. *Journal of Experimental Medicine*, 209(10):1769–1779, sep 2012. doi: 10.1084/jem.20120905.

Sonia Feau, Zacarias Garcia, Ramon Arens, Hideo Yagita, Jannie Borst, and Stephen P. Schoenberger. The CD4+ T-cell help signal is transmitted from APC to CD8+ T-cells via CD27-CD70 interactions. *Nature Communications*, 3(1), jan 2012. doi: 10.1038/ncomms1948.

Olivier Fisette, Sebastian Wingbermühle, Robert Tampé, and Lars V. Schäfer. Molecular mechanism of peptide editing in the tapasin–MHC I complex. *Scientific Reports*, 6(1), jan 2016. doi: 10.1038/srep19085.

B J Fowlkes, L Edison, B J Mathieson, and T M Chused. Early T lymphocytes. differentiation in vivo of adult intrathymic precursor cells. *Journal of Experimental Medicine*, 162(3):802–822, sep 1985. doi: 10.1084/jem.162.3.802.

Jonah Gabry, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. Visualization in bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(2):389–402, jan 2019. doi: 10.1111/rssa.12378.

Andrew Gelman. The folk theorem of statistical computing, 2008. URL `https://statmodeling.stat.columbia.edu/2008/05/13/the_folk_theore/`. Accessed on 23.03.2022.

Andrew Gelman. Commentary. *Epidemiology*, 24(1):69–72, jan 2013. doi: 10.1097/ede.0b013e31827886f7.

Andrew Gelman and John Carlin. Beyond power calculations. *Perspectives on Psychological Science*, 9(6):641–651, nov 2014. doi: 10.1177/1745691614551642.

Andrew Gelman and Christian Hennig. Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(4): 967–1033, aug 2017. doi: 10.1111/rssa.12276.

Andrew Gelman and Iain Pardoe. Average predictive comparisons for models with nonlinearity, interactions, and variance components. *Sociological Methodology*, 37:23–51, 2007. ISSN 00811750, 14679531. URL `http://www.jstor.org/stable/20451130`.

Andrew Gelman and Hal Stern. The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, 60(4):328–331, nov 2006. doi: 10.1198/000313006x152649.

Andrew Gelman, Daniel Simpson, and Michael Betancourt. The prior can often only be understood in the context of the likelihood. *Entropy*, 19(10):555, oct 2017. doi: 10.3390/e19100555.

Andrew Gelman, Aki Vehtari, Daniel Simpson, Charles C Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. Bayesian workflow. *arXiv preprint arXiv:2011.01808*, 2020.

D I Godfrey, J Kennedy, T Suda, and A Zlotnik. A developmental pathway involving four phenotypically and functionally distinct subsets of CD3-CD4-CD8- triple-negative adult mouse thymocytes defined by CD44 and CD25 expression. *The Journal of Immunology*, 150(10):4244–4452, may 1993.

Ben Goodrich, Jonah Gabry, Imad Ali, and Sam Brilleman. rstanarm: Bayesian applied regression modeling via Stan., 2020. URL `https://mc-stan.org/rstanarm`. R package version 2.21.1.

Jörg J Goronzy and Cornelia M Weyand. T-cell co-stimulatory pathways in autoimmunity. *Arthritis Research & Therapy*, 10(Suppl 1):S3, 2008. doi: 10.1186/ar2414.

Klaus Gossens, Silvia Naus, Stephane Y. Corbel, Shujun Lin, Fabio M.V. Rossi, Jürgen Kast, and Hermann J. Ziltener. Thymic progenitor homing and lymphocyte homeostasis are linked via S1P-controlled expression of thymic P-selectin/CCL25. *Journal of Experimental Medicine*, 206(4):761–778, mar 2009. doi: 10.1084/jem.20082502.

Sander Greenland. Valid p-values behave exactly as they should: Some misleading criticisms of p-values and their resolution with s-values. *The American Statistician*, 73 (sup1):106–114, mar 2019. doi: 10.1080/00031305.2018.1529625.

H. M. Grey, J. Ruppert, A. Vitiello, J. Sidney, W. M. Kast, R. T. Kubo, and A. Sette. Class I MHC-peptide interactions: structural requirements and functional implications. *Cancer Surveys*, 22:37–49, 1995. ISSN 0261-2429.

Guinevere L. Grice and James A. Nathan. The recognition of ubiquitinated proteins by the proteasome. *Cellular and Molecular Life Sciences*, 73(18):3497–3506, may 2016. doi: 10.1007/s00018-016-2255-5.

Michael Groll, Lars Ditzel, Jan Löwe, Daniela Stock, Matthias Bochtler, Hans D. Bartunik, and Robert Huber. Structure of 20S proteasome from yeast at 2.4Å resolution. *Nature*, 386(6624):463–471, April 1997. doi: 10.1038/386463a0.

T Morgan H Sepulveda, A Cerwenka and RW Dutton. CD28, IL-2-independent costimulatory pathways for CD8 T lymphocyte activation. *The Journal of Immunology*, 150 (10):4244–4452, may 1999.

Daniel Habermann, Hadi Kharimzadeh, Andreas Walker, Yang Li, Rongge Yang, Rolf Kaiser, Zabrina L Brumme, Jörg Timm, Michael Roggendorf, and Daniel Hoffmann. HAMdetector: a Bayesian regression model that integrates information to detect HLA-associated mutations. *Bioinformatics*, 03 2022. ISSN 1367-4803. doi: 10.1093/bioinformatics/btac134. URL `https://doi.org/10.1093/bioinformatics/btac134`. btac134.

Q. Han, N. Bagheri, E. M. Bradshaw, D. A. Hafler, D. A. Lauffenburger, and J. C. Love. Polyfunctional responses by human T cells result from sequential release of cytokines. *Proceedings of the National Academy of Sciences*, 109(5):1607–1612, dec 2011. doi: 10.1073/pnas.1117194109.

Kornelia Heinzel, Claudia Benz, Vera C. Martins, Ian D. Haidl, and Conrad C. Bleul. Bone marrow-derived hemopoietic precursors commit to the T cell lineage only after arrival in the thymic microenvironment. *The Journal of Immunology*, 178(2):858–868, jan 2007. doi: 10.4049/jimmunol.178.2.858.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. ISSN 00401706. URL `http://www.jstor.org/stable/1267351`.

Victor W. Hsu, Lydia C. Yuan, Jed G. Nuchtern, Jennifer Lippincott-Schwartz, Gunter J. Hammerling, and Richard D. Klausner. A recycling pathway between the endoplasmic reticulum and the golgi apparatus for retention of unassembled MHC class I molecules. *Nature*, 352(6334):441–444, aug 1991. doi: 10.1038/352441a0.

Ching-Yu Huang and Osami Kanagawa. Ordered and coordinated rearrangement of the TCR α locus: Role of secondary rearrangement in thymic selection. *The Journal of Immunology*, 166(4):2597–2601, feb 2001. doi: 10.4049/jimmunol.166.4.2597.

Koraljka Husnjak, Suzanne Elsasser, Naixia Zhang, Xiang Chen, Leah Randles, Yuan Shi, Kay Hofmann, Kylie J. Walters, Daniel Finley, and Ivan Dikic. Proteasome subunit Rpn13 is a novel ubiquitin receptor. *Nature*, 453(7194):481–488, may 2008. doi: 10.1038/nature06926.

Masaaki Ibe, Yuki Ikeda Moore, Kiyoshi Miwa, Yutaro Kaneko, Shumpei Yokota, and Masafumi Takiguchi. Role of strong anchor residues in the effective binding of 10-mer and 11-mer peptides to HLA-A*2402 molecules. *Immunogenetics*, 44(4):233–241, jul 1996. doi: 10.1007/bf02602551.

John P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2 (8):e124, aug 2005. doi: 10.1371/journal.pmed.0020124.

Sheena S. Iyengar and Mark R. Lepper. When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology*, 79(6):995–1006, dec 2000. doi: 10.1037/0022-3514.79.6.995.

Sir Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461, sep 1946. doi: 10.1098/rspa.1946.0056.

D Kagi, F Vignaux, B Ledermann, K Burki, V Depraetere, S Nagata, H Hengartner, and P Golstein. Fas and perforin pathways as major mechanisms of T cell-mediated cytotoxicity. *Science*, 265(5171):528–530, jul 1994. doi: 10.1126/science.7518614.

Alexei F. Kisselev, Tatos N. Akopian, Kee Min Woo, and Alfred L. Goldberg. The sizes of peptides generated from protein by mammalian 26 and 20 S proteasomes. *Journal of Biological Chemistry*, 274(6):3363–3371, feb 1999. doi: 10.1074/jbc.274.6.3363.

Jan Klein, Hideki Ono, Dagmar Klein, and Colm O'hUigin. The accordion model of MHC evolution. In *Progress in Immunology Vol. VIII*, pages 137–143. Springer Berlin Heidelberg, 1993. doi: 10.1007/978-3-642-51479-1_18.

Ludger Klein, Bruno Kyewski, Paul M. Allen, and Kristin A. Hogquist. Positive and negative selection of the t cell repertoire: what thymocytes see (and don't see). *Nature Reviews Immunology*, 14(6):377–391, may 2014. doi: 10.1038/nri3667.

Andrei N Kolmogorov. Foundations of the theory of probability, 2nd english edition. *NY: Chelsea Publishing Co*, 1933.

Jens-Oliver Koopmann, Markus Post, Jacques J. Neefjes, Günter J. Hämmerling, and Frank Momburg. Translocation of long peptides by transporters associated with antigen processing (TAP). *European Journal of Immunology*, 26(8):1720–1728, aug 1996. doi: 10.1002/eji.1830260809.

John K. Kruschke. Optional stopping in data collection: p values, bayes factors, credible intervals, precision, 2013. URL `http://doingbayesiandataanalysis.blogspot.com/2013/11/optional-stopping-in-data-collection-p.html`. Accessed on 14.03.2022.

Brahma V. Kumar, Thomas J. Connors, and Donna L. Farber. Human T cell development, localization, and function throughout life. *Immunity*, 48(2):202–213, feb 2018. doi: 10.1016/j.immuni.2018.01.007.

Alexandr Kuznetsov, Alice Voronina, Vadim Govorun, and Georgij Arapidi. Critical review of existing MHC I immunopeptidome isolation methods. *Molecules*, 25(22): 5409, nov 2020. doi: 10.3390/molecules25225409.

Joël Lanoix, Chantal Durette, Mathieu Courcelles, Émilie Cossette, Simon Comtois-Marotte, Marie-Pierre Hardy, Caroline Côté, Claude Perreault, and Pierre Thibault. Comparison of the MHC I immunopeptidome repertoire of B-cell lymphoblasts using two isolation methods. *PROTEOMICS*, 18(12):1700251, apr 2018. doi: 10.1002/pmic.201700251.

Pierre Simon Laplace. *A philosophical essay on probabilities.* Dover Publications Inc., 1814.

A. Lew, D. Pardoll, W. Maloy, B. Fowlkes, A Kruisbeek, S. Cheng, R. Germain, J. Bluestone, R. Schwartz, and J. Coligan. Characterization of T cell receptor gamma chain expression in a subset of murine thymocytes. *Science*, 234(4782):1401–1405, dec 1986. doi: 10.1126/science.3787252.

Lenong Li, Mansoor Batliwala, and Marlene Bouvier. ERAP1 enzyme-mediated trimming and structural analyses of MHC i–bound precursor peptides yield novel insights into antigen processing and presentation. *Journal of Biological Chemistry*, 294(49): 18534–18544, dec 2019. doi: 10.1074/jbc.ra119.010102.

Ling Li, Mei Dong, and Xiao-Guang Wang. The implication and significance of beta 2 microglobulin. *Chinese Medical Journal*, 129(4):448–455, feb 2016. doi: 10.4103/0366-6999.176084.

Yili Li, Yiyuan Yin, and Roy A. Mariuzza. Structural and biophysical insights into the role of CD4 and CD8 in T cell activation. *Frontiers in Immunology*, 4, 2013. doi: 10.3389/fimmu.2013.00206.

Roland S. Liblau, F.Susan Wong, Lennart T. Mars, and Pere Santamaria. Autoreactive CD8 T cells in organ-specific autoimmunity. *Immunity*, 17(1):1–6, jul 2002. doi: 10.1016/s1074-7613(02)00338-2.

Juliane Liepe, Fabio Marino, John Sidney, Anita Jeko, Daniel E. Bunting, Alessandro Sette, Peter M. Kloetzel, Michael P. H. Stumpf, Albert J. R. Heck, and Michele Mishto. A large fraction of HLA class I ligands are proteasome-generated spliced peptides. *Science*, 354(6310):354–358, oct 2016. doi: 10.1126/science.aaf4384.

Rishi Vishal Luckheeram, Rui Zhou, Asha Devi Verma, and Bing Xia. CD4+ T cells: Differentiation and functions. *Clinical and Developmental Immunology*, 2012:1–12, 2012. doi: 10.1155/2012/925135.

Judith N. Mandl, João P. Monteiro, Nienke Vrisekoop, and Ronald N. Germain. T cell-positive selection uses self-ligand binding strength to optimize repertoire recognition of foreign antigens. *Immunity*, 38(2):263–274, feb 2013. doi: 10.1016/j.immuni.2012.09.011.

S. G. E. Marsh, E. D. Albert, W. F. Bodmer, R. E. Bontrop, B. Dupont, H. A. Erlich, M. Fernández-Viña, D. E. Geraghty, R. Holdsworth, C. K. Hurley, M. Lau, K. W. Lee, B. Mach, M. Maiers, W. R. Mayr, C. R. Müller, P. Parham, E. W. Petersdorf, T. Sasazuki, J. L. Strominger, A. Svejgaard, P. I. Terasaki, J. M. Tiercy, and J. Trowsdale. Nomenclature for factors of the HLA system, 2010. *Tissue Antigens*, 75(4): 291–455, apr 2010. doi: 10.1111/j.1399-0039.2010.01466.x.

Andreas Martin, Tania A Baker, and Robert T Sauer. Pore loops of the AAA+ ClpX machine grip substrates to drive translocation and unfolding. *Nature Structural & Molecular Biology*, 15(11):1147–1151, oct 2008. doi: 10.1038/nsmb.1503.

Javier Mestas and Christopher C. W. Hughes. Of mice and not men: Differences between mouse and human immunology. *The Journal of Immunology*, 172(5):2731–2738, feb 2004. doi: 10.4049/jimmunol.172.5.2731.

Booki Min, Rebecca McHugh, Gregory D Sempowski, Crystal Mackall, Gilles Foucras, and William E Paul. Neonates support lymphopenia-induced proliferation. *Immunity*, 18(1):131–140, jan 2003. doi: 10.1016/s1074-7613(02)00508-3.

Sebastian Montealegre and Peter M. van Endert. Endocytic recycling of MHC class I molecules in non-professional antigen presenting and dendritic cells. *Frontiers in Immunology*, 9, jan 2019. doi: 10.3389/fimmu.2018.03098.

D L Mueller, M K Jenkins, and R H Schwartz. Clonal expansion versus functional clonal inactivation: A costimulatory signalling pathway determines the outcome of T cell antigen receptor occupancy. *Annual Review of Immunology*, 7(1):445–480, apr 1989. doi: 10.1146/annurev.iy.07.040189.002305.

Kenneth Murphy. *Janeway's immunobiology*. Garland Science, New York, 2017. ISBN 978-0815345053.

Tomas Mustelin and Kjetil Taskén. Positive and negative regulation of T-cell activation through kinases and phosphatases. *Biochemical Journal*, 371(1):15–27, apr 2003. doi: 10.1042/bj20021637.

J. Neefjes, F Momburg, and G. Hammerling. Selective and ATP-dependent translocation of peptides by the MHC-encoded transporter. *Science*, 261(5122):769–771, aug 1993. doi: 10.1126/science.8342042.

Masatoshi Nei and Alejandro P. Rooney. Concerted and birth-and-death evolution of multigene families. *Annual Review of Genetics*, 39(1):121–152, dec 2005. doi: 10.1146/annurev.genet.39.073003.112240.

Jerzy Neyman and Egon Sharpe Pearson. IX. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706): 289–337, feb 1933. doi: 10.1098/rsta.1933.0009.

Janko Nikolić-Žugić. Phenotypic and functional stages in the intrathymic development of $\alpha\beta$ T cells. *Immunology Today*, 12(2):65–70, feb 1991. doi: 10.1016/0167-5699(91)90160-u.

Timothy J. O'Donnell, Alex Rubinsteyn, and Uri Laserson. MHCflurry 2.0: Improved pan-allele prediction of MHC Class I-presented peptides by incorporating antigen processing. *Cell Systems*, 11(1):42–48.e7, jul 2020. doi: 10.1016/j.cels.2020.06.010.

Gail Pearse. Normal structure, function and histology of the thymus. *Toxicologic Pathology*, 34(5):504–514, aug 2006. doi: 10.1080/01926230600865549.

H T Petrie, F Livak, D G Schatz, A Strasser, I N Crispe, and K Shortman. Multiple rearrangements in T cell receptor alpha chain genes maximize the production of useful thymocytes. *Journal of Experimental Medicine*, 178(2):615–622, aug 1993. doi: 10.1084/jem.178.2.615.

Stuart J. Pocock. *Clinical Trials - A Practical Approach*. Wiley, New York, 1983. ISBN 978-0-471-90155-6.

Michael F. Princiotta, Diana Finzi, Shu-Bing Qian, James Gibbs, Sebastian Schuchmann, Frank Buttgereit, Jack R. Bennink, and Jonathan W. Yewdell. Quantitating protein synthesis, degradation, and endogenous antigen processing. *Immunity*, 18(3):343–354, mar 2003. doi: 10.1016/s1074-7613(03)00051-7.

John C Pui, David Allman, Lanwei Xu, Susan DeRocco, Fredrick G Karnell, Sonia Bakkour, Julia Y Lee, Tom Kadesch, Richard R Hardy, Jon C Aster, and Warren S Pear. Notch1 expression in early lymphopoiesis influences B versus T lineage determination. *Immunity*, 11(3):299–308, sep 1999. doi: 10.1016/s1074-7613(00)80105-3.

H G Rammensee, K Falk, and O Rötzschke. Peptides naturally presented by MHC class I molecules. *Annual Review of Immunology*, 11(1):213–244, apr 1993a. doi: 10.1146/annurev.iy.11.040193.001241.

Hans-Georg Rammensee, Kirsten Falk, and Olaf Rötzschke. MHC molecules as peptide receptors. *Current Opinion in Immunology*, 5(1):35–44, feb 1993b. doi: 10.1016/0952-7915(93)90078-7.

Pedro A. Reche and Ellis L. Reinherz. Sequence variability analysis of human class I and class II MHC molecules: Functional and structural correlates of amino acid polymorphisms. *Journal of Molecular Biology*, 331(3):623–641, aug 2003. doi: 10.1016/s0022-2836(03)00750-2.

Eric Reits, Joost Neijssen, Carla Herberts, Willemien Benckhuijsen, Lennert Janssen, Jan Wouter Drijfhout, and Jacques Neefjes. A major role for TPPII in trimming proteasomal degradation products for MHC class I antigen presentation. *Immunity*, 20 (4):495–506, apr 2004. doi: 10.1016/s1074-7613(04)00074-3.

Birkir Reynisson, Bruno Alvarez, Sinu Paul, Bjoern Peters, and Morten Nielsen. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Research*, 48(W1):W449–W454, may 2020. doi: 10.1093/nar/gkaa379.

Melissa J. Rist, Alex Theodossis, Nathan P. Croft, Michelle A. Neller, Andrew Welland, Zhenjun Chen, Lucy C. Sullivan, Jacqueline M. Burrows, John J. Miles, Rebekah M. Brennan, Stephanie Gras, Rajiv Khanna, Andrew G. Brooks, James McCluskey, Anthony W. Purcell, Jamie Rossjohn, and Scott R. Burrows. HLA peptide length preferences control CD8+ T cell responses. *The Journal of Immunology*, 191(2):561–571, jun 2013. doi: 10.4049/jimmunol.1300292.

J. Robinson. OPTIMAL TESTS OF SIGNIFICANCE. *Australian Journal of Statistics*, 21(3):301–310, sep 1979. doi: 10.1111/j.1467-842x.1979.tb01147.x.

James Robinson, Lisbeth A. Guethlein, Nezih Cereb, Soo Young Yang, Paul J. Norman, Steven G. E. Marsh, and Peter Parham. Distinguishing functional polymorphism from random variation in the sequences of >10,000 HLA-A, -B and -C alleles. *PLOS Genetics*, 13(6):e1006862, jun 2017. doi: 10.1371/journal.pgen.1006862.

James Robinson, Dominic J Barker, Xenia Georgiou, Michael A Cooper, Paul Flicek, and Steven G E Marsh. IPD-IMGT/HLA database. *Nucleic Acids Research*, oct 2019. doi: 10.1093/nar/gkz950.

Justine Rochon, Matthias Gondan, and Meinhard Kieser. To test or not to test: Preliminary assessment of normality when comparing two independent samples. *BMC Medical Research Methodology*, 12(1), jun 2012. doi: 10.1186/1471-2288-12-81.

Sarah H. Ross and Doreen A. Cantrell. Signaling and function of Interleukin-2 in T lymphocytes. *Annual Review of Immunology*, 36(1):411–433, apr 2018. doi: 10.1146/annurev-immunol-042617-053352.

Sean O. Ryan and Brian A. Cobb. Host glycans and antigen presentation. *Microbes and Infection*, 14(11):894–903, sep 2012. doi: 10.1016/j.micinf.2012.04.010.

Lawrence E. Samelson, Maitray D. Patel, Allan M. Weissman, Joe B. Harford, and Richard D. Klausner. Antigen activation of murine T cells induces tyrosine phosphorylation of a polypeptide associated with the T cell antigen receptor. *Cell*, 46(7): 1083–1090, sep 1986. doi: 10.1016/0092-8674(86)90708-7.

Fadil Santosa and William W. Symes. Linear inversion of band-limited reflection seismograms. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1307–1330, oct 1986. doi: 10.1137/0907087.

Leonard J Savage. *The foundations of statistics.* Courier Corporation, 1972.

Patricia Saxová, Søren Buus, Søren Brunak, and Can Keşmir. Predicting proteasomal cleavage sites: a comparison of available methods. *International Immunology*, 15(7): 781–787, jul 2003. doi: 10.1093/intimm/dxg084.

D G Schatz, M A Oettinger, and M S Schlissel. V(D)Jrecombination: Molecular biology and regulation. *Annual Review of Immunology*, 10(1):359–383, apr 1992. doi: 10.1146/annurev.iy.10.040192.002043.

Heiko Schuster, Janet K. Peper, Hans-Christian Bösmüller, Kevin Röhle, Linus Backert, Tatjana Bilich, Britta Ney, Markus W. Löffler, Daniel J. Kowalewski, Nico Trautwein, Armin Rabsteyn, Tobias Engler, Sabine Braun, Sebastian P. Haen, Juliane S. Walz, Barbara Schmid-Horch, Sara Y. Brucker, Diethelm Wallwiener, Oliver Kohlbacher, Falko Fend, Hans-Georg Rammensee, Stefan Stevanović, Annette Staebler, and Philipp Wagner. The immunopeptidomic landscape of ovarian carcinomas. *Proceedings of the National Academy of Sciences*, 114(46):E9942–E9951, nov 2017. doi: 10.1073/pnas.1707658114.

Ronald H. Schwartz. T cell anergy. *Annual Review of Immunology*, 21(1):305–334, apr 2003. doi: 10.1146/annurev.immunol.21.120601.141110.

Thomas Serwold, Lauren I. Richie Ehrlich, and Irving L. Weissman. Reductive isolation from bone marrow and blood implicates common lymphoid progenitors as the major source of thymopoiesis. *Blood*, 113(4):807–815, jan 2009. doi: 10.1182/blood-2008-08-173682.

James C. Shepherd, Ton N.M. Schumacher, Philip G. Ashton-Rickardt, Suguru Imaeda, Hidde L. Ploegh, Charles A. Janeway, and Susumu Tonegawa. TAP1-dependent peptide translocation in vitro is ATP dependent and peptide selective. *Cell*, 74(3):577–584, aug 1993. doi: 10.1016/0092-8674(93)80058-m.

Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. False-positive psychology. *Psychological Science*, 22(11):1359–1366, oct 2011. doi: 10.1177/0956797611417632.

Jennifer E. Smith-Garvin, Gary A. Koretzky, and Martha S. Jordan. T cell activation. *Annual Review of Immunology*, 27(1):591–619, apr 2009. doi: 10.1146/annurev.immunol.021908.132706.

Hae Won Sohn, Young Kee Shin, Im-Soon Lee, Young Mee Bae, Young Ho Suh, Min Kyung Kim, Tae Jin Kim, Kyeong Cheon Jung, Weon Seo Park, Chan-Sik Park, Doo Hyun Chung, Kwangseog Ahn, In Sun Kim, Young Hyeh Ko, Yung Jue Bang, Chul Woo Kim, and Seong Hoe Park. CD99 regulates the transport of MHC class I molecules from the golgi complex to the cell surface. *The Journal of Immunology*, 166 (2):787–794, jan 2001. doi: 10.4049/jimmunol.166.2.787.

Ivelina Spassova, Selma Ugurel, Patrick Terheyden, Antje Sucker, Jessica C. Hassel, Cathrin Ritter, Linda Kubat, Daniel Habermann, Farnoush Farahpour, Mohammadkarim Saeedghalati, Lukas Peiffer, Rajiv Kumar, David Schrama, Daniel Hoffmann, Dirk Schadendorf, and Jürgen C. Becker. Predominance of central memory t cells with high t-cell receptor repertoire diversity is associated with response to PD-1/PD-l1 inhibition in merkel cell carcinoma. *Clinical Cancer Research*, 26(9):2257–2267, jan 2020. doi: 10.1158/1078-0432.ccr-19-2244.

Ivelina Spassova, Selma Ugurel, Linda Kubat, Lisa Zimmer, Patrick Terheyden, Annalena Mohr, Hannah Björn Andtback, Lisa Villabona, Ulrike Leiter, Thomas Eigentler, Carmen Loquai, Jessica C Hassel, Thilo Gambichler, Sebastian Haferkamp, Peter Mohr, Claudia Pfoehler, Lucie Heinzerling, Ralf Gutzmer, Jochen S Utikal, Kai Horny, Hans-Ulrich Schildhaus, Daniel Habermann, Daniel Hoffmann, Dirk Schadendorf, and Jürgen Christian Becker. Clinical and molecular characteristics associated with response to therapeutic PD-1/PD-l1 inhibition in advanced merkel cell carcinoma. *Journal for ImmunoTherapy of Cancer*, 10(1):e003198, jan 2022. doi: 10.1136/jitc-2021-003198.

Anette Stryhn, Lars Østergaard Pedersen, Arne Holm, and Søren Buus. Longer peptide can be accommodated in the MHC class I binding site by a protrusion mechanism. *European Journal of Immunology*, 30(11):3089–3099, nov 2000. doi: 10.1002/1521-4141(200011)30:11<3089::aid-immu3089>3.0.co;2-5.

Charles D. Surh and Jonathan Sprent. T-cell apoptosis detected in situ during positive and negative selection in the thymus. *Nature*, 372(6501):100–103, nov 1994. doi: 10.1038/372100a0.

Hiroyuki Takaba and Hiroshi Takayanagi. The mechanisms of T cell selection in the thymus. *Trends in Immunology*, 38(11):805–816, nov 2017. doi: 10.1016/j.it.2017.07.010.

The MHC sequencing consortium. Complete sequence and gene map of a human major histocompatibility complex. *Nature*, 401(6756):921–923, oct 1999. doi: 10.1038/44853.

E. Thorsby. A short history of HLA. *Tissue Antigens*, 74(2):101–116, aug 2009. doi: 10.1111/j.1399-0039.2009.01291.x.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246. URL http://www.jstor.org/stable/2346178.

Joseph A. Trapani and Mark J. Smyth. Functional significance of the perforin/granzyme cell death pathway. *Nature Reviews Immunology*, 2(10):735–747, oct 2002. doi: 10.1038/nri911.

S. Uebel, W. Kraas, S. Kienle, K.-H. Wiesmuller, G. Jung, and R. Tampe. Recognition principle of the TAP transporter disclosed by combinatorial peptide libraries. *Proceedings of the National Academy of Sciences*, 94(17):8976–8981, aug 1997. doi: 10.1073/pnas.94.17.8976.

Ramunas M Vabulas. Proteasome function and protein biosynthesis. *Current Opinion in Clinical Nutrition and Metabolic Care*, 10(1):24–31, jan 2007. doi: 10.1097/mco.0b013e328011645b.

Peter M. van Endert, Robert Tampé, Thomas H. Meyer, Roland Tisch, Jean-François Bach, and Hugh O. McDevitt. A sequential model for peptide binding and transport by the transporters associated with antigen processing. *Immunity*, 1(6):491–500, sep 1994. doi: 10.1016/1074-7613(94)90091-4.

John Venn. *The logic of chance: an essay on the foundations and province of the theory of probability, with especial reference to its logical bearings and its application to moral and social science, and to statistics*. Macmillan, 1888.

Harald von Boehmer and Hans Jörg Fehling. STRUCTURE AND FUNCTION OF THE PRE-T CELL RECEPTOR. *Annual Review of Immunology*, 15(1):433–452, apr 1997. doi: 10.1146/annurev.immunol.15.1.433.

Matthew A. Williams and Michael J. Bevan. Effector and memory CTL differentiation. *Annual Review of Immunology*, 25(1):171–192, apr 2007. doi: 10.1146/annurev.immunol.25.022106.141548.

Margreet A Wolfert and Geert-Jan Boons. Adaptive immune activation: glycosylation does matter. *Nature Chemical Biology*, 9(12):776–784, nov 2013. doi: 10.1038/nchembio.1403.

Sho Yamasaki, Eri Ishikawa, Machie Sakuma, Koji Ogata, Kumiko Sakata-Sogawa, Michio Hiroshima, David L Wiest, Makio Tokunaga, and Takashi Saito. Mechanistic basis of pre–T cell receptor–mediated autonomous signaling critical for thymocyte development. *Nature Immunology*, 7(1):67–75, dec 2005. doi: 10.1038/ni1290.

Daniel A. Zlotoff and Avinash Bhandoola. Hematopoietic progenitor migration to the adult thymus. *Annals of the New York Academy of Sciences*, 1217(1):122–138, jan 2011. doi: 10.1111/j.1749-6632.2010.05881.x.

P. Zwickl, D. Voges, and W. Baumeister. The proteasome: a macromolecular assembly designed for controlled proteolysis. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 354(1389):1501–1511, sep 1999. doi: 10.1098/rstb.1999.0494.

Appendix A

List of Figures, Tables and Abbreviations

List of Figures

List of Tables

List of Abbreviations

ATP Adenosine triphosphate

CV Cross-validation

ER Endoplasmatic reticulum

HAM HLA-associated mutation

HBV Hepatitis B Virus

HCV Hepatitis C Virus

HDV Hepatitis D Virus

HIV Human immunodeficiency virus

HLA Human Leukocyte Antigen

LOO Leave-one-out

MCC Merkel cell carcinoma

MHC Major histocompatibility complex

NHST Null hypothesis significance test

PLC Peptide loading complex

TAP Transporter for antigen processing

TCR T cell receptor

Appendix B

Glossary of Probability Distributions

Continuous distributions

Table B.1: List of continuous probability distributions, with their probability density functions and parameters; exp is the exponential function; $\mathbb{R}$ set of real numbers; $\mathbb{R}^+$ set of positive real numbers excluding 0; $\mathbb{R}_0^+$ set of positive real numbers including 0. Γ is the Gamma function $\Gamma(x) = \int_0^\infty u^{x-1}\exp(-u)du$; B is the Beta function $B(u,v) = \dfrac{\Gamma(u)\Gamma(v)}{\Gamma(u)+\Gamma(v)}$ for $u \in \mathbb{R}^+$ and $v \in \mathbb{R}^+$.

Distribution	Notation and pdf	Data and parameters
Normal	$p(y) = \text{Normal}(y\|\mu,\sigma)$ $= \dfrac{1}{\sqrt{2\pi}\sigma}\exp\left(-\dfrac{1}{2\sigma^2}(y-\mu)^2\right)$	$y \in \mathbb{R}$, $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$
Gamma	$p(y) = \text{Gamma}(y\|\alpha,\beta)$ $= \dfrac{\beta^\alpha}{\Gamma(\alpha)}y^{\alpha-1}\exp(-\beta y)$	$y \in \mathbb{R}^+, \alpha \in \mathbb{R}^+, \beta \in \mathbb{R}^+$
Exponential	$p(y) = \text{Exponential}(y\|\beta)$ $= \beta\exp(-\beta y)$	$y \in \mathbb{R}_0^+, \beta \in \mathbb{R}^+$
Cauchy	$p(y) = \text{Cauchy}(y\|\mu,\sigma)$ $= \dfrac{1}{\pi\sigma}\dfrac{1}{1+((y-\mu)/\sigma^2)}$	$y \in \mathbb{R}, \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+$
Beta	$p(y) = \text{Beta}(y\|\alpha,\beta)$ $= \dfrac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha,\beta)}$	$y \in [0,1], \alpha \in \mathbb{R}^+, \beta \in \mathbb{R}^+$

Discrete distributions

Table B.2: List of discrete probability distributions, with their probability mass functions and parameters; $\mathbb{N}$ set of natural numbers including 0; $\mathbb{R}$ set of real numbers; $\mathbb{R}^+$ set of positive real numbers excluding 0; Γ is the Gamma function $\Gamma(x) = \int_0^\infty u^{x-1} \exp(-u)du$; B is the Beta function $B(u,v) = \dfrac{\Gamma(u)\Gamma(v)}{\Gamma(u)+\Gamma(v)}$ for $u \in \mathbb{R}^+$ and $v \in \mathbb{R}^+$.

Distribution	Notation and pmf	Data and parameters
Poisson	$p(y) = \text{Poisson}(y\|\lambda)$ $= \dfrac{1}{y!}\lambda^y \exp(-\lambda)$	$y \in \mathbb{N}, \lambda \in \mathbb{R}^+$
Binomial	$p(y) = \text{Binomial}(y\|n,\theta)$ $= \binom{n}{y}\theta^y \left(1-\theta\right)^{n-y}$	$y \in 0,\ldots,n, \theta \in [0,1], n \in \mathbb{N}$
Beta-Binomial	$p(y) = \text{Beta-Binomial}(y\|n,\alpha,\beta)$ $= \binom{n}{y}\dfrac{B(y+\alpha, n-y+\beta)}{B(\alpha,\beta)}$	$y \in 0,\ldots,n, \alpha \in \mathbb{R}^+, \beta \in \mathbb{R}^+, n \in \mathbb{N}$

Declaration of Contribution

Kumulative Dissertation/Beteiligung an Veröffentlichungen

Kumulative Dissertation von Herrn Daniel Habermann

Autorenbeiträge

Titel der Publikation: HAMdetector: A Bayesian regression model that integrates information to detect HLA-associated mutations

Autoren: Daniel Habermann, Hadi Kharimzadeh, Andreas Walker, Yang Li, Rongge Yang, Rolf Kaiser, Zabrina L. Brumme, Jörg Timm, Michael Roggendorf, Daniel Hoffmann

Anteile:

- Konzept - 75%
- Durchführung der Experimente - 0%
- Datenanalyse - 100%
- Artenanalyse - NA%
- Statistische Analyse - 100%
- Manuskripterstellung - 50%
- Überarbeitung des Manuskripts - 50%

Unterschrift Doktorand/in

Unterschrift Betreuer/in

Kumulative Dissertation/Beteiligung an Veröffentlichungen

Kumulative Dissertation von Herrn Daniel Habermann

Autorenbeiträge

Titel der Publikation: Clinical and molecular characteristics associated with response to therapeutic PD-1/PD-L1 inhibition in advanced Merkel cell carcinoma

Autoren: Ivelina Spassova, Selma Ugurel, Linda Kubat, Lisa Zimmer, Patrick Terheyden, Annalena Mohr, Hannah Björn Andtback, Lisa Villabona, Ulrike Leiter, Thomas Eigentler, Carmen Loquai, Jessica C. Hassel, Thilo Gambichler, Sebastian Haferkamp, Peter Mohr, Claudia Pfoehler, Lucie Heinzerling, Ralf Gutzmer, Jochen S Utikal, Kai Horny, Hans-Ulrich Schildhaus, Daniel Habermann, Daniel Hoffmann, Dirk Schadendorf, Jürgen Christian Becker

Anteile:

- Konzept - 0%
- Durchführung der Experimente - 0%
- Datenanalyse - 50%
- Artenanalyse - NA%
- Statistische Analyse - 75%
- Manuskripterstellung - 10%
- Überarbeitung des Manuskripts - 10%

Unterschrift Doktorand/in

Unterschrift Betreuer/in

Kumulative Dissertation/Beteiligung an Veröffentlichungen

Kumulative Dissertation von Herrn Daniel Habermann

Autorenbeiträge

Titel der Publikation: Predominance of Central Memory T Cells with High T-Cell Receptor Repertoire Diversity is Associated with Response to PD-1/PD-L1 Inhibition in Merkel Cell Carcinoma

Autoren: Ivelina Spassova, Selma Ugurel, Patrick Terheyden, Antje Sucker, Jessica C Hassel, Cathrin Ritter, Linda Kubat, Daniel Habermann, Farnoush Farahpour, Mohammadkarim Saeedghalati, Lukas Pfeiffer, Rajiv Kumar, David Schrama, Daniel Hoffmann, Dirk Schadendorf, Jürgen C. Becker

Anteile:

- Konzept - 0%
- Durchführung der Experimente - 0%
- Datenanalyse - 50%
- Artenanalyse - NA%
- Statistische Analyse - 75%
- Manuskripterstellung - 10%
- Überarbeitung des Manuskripts - 10%

Unterschrift Doktorand/in

Unterschrift Betreuer/in

The curriculum vitae is not included in the online version for data protection reasons.

The curriculum vitae is not included in the online version for data protection reasons.

Declarations

Erklärung:
Hiermit erkläre ich, gem. § 7 Abs. (2) d) + f) der Promotionsordnung der Fakultät für Biologie zur Erlangung des Dr. rer. nat., dass ich die vorliegende Dissertation selbständig verfasst und mich keiner anderen als der angegebenen Hilfsmittel bedient, bei der Abfassung der Dissertation nur die angegeben Hilfsmittel benutzt und alle wörtlich oder inhaltlich übernommenen Stellen als solche gekennzeichnet habe.

Essen, den ________________ ____

Unterschrift des/r Doktoranden/in

Erklärung:
Hiermit erkläre ich, gem. § 7 Abs. (2) e) + g) der Promotionsordnung der Fakultät für Biologie zur Erlangung des Dr. rer. nat., dass ich keine anderen Promotionen bzw. Promotionsversuche in der Vergangenheit durchgeführt habe und dass diese Arbeit von keiner anderen Fakultät/Fachbereich abgelehnt worden ist.

Essen, den ________________ ____

Unterschrift des/r Doktoranden/in

Erklärung:
Hiermit erkläre ich, gem. § 6 Abs. (2) g) der Promotionsordnung der Fakultät für Biologie zur Erlangung der Dr. rer. nat., dass ich das Arbeitsgebiet, dem das Thema "Computational methods to detect HLA-associated mutations" zuzuordnen ist, in Forschung und Lehre vertrete und den Antrag von Daniel Habermann befürworte und die Betreuung auch im Falle eines Weggangs, wenn nicht wichtige Gründe dem entgegenstehen, weiterführen werde.

____

Name des Mitglieds der Universität Duisburg-Essen in Druckbuchstaben

Essen, den ____

Unterschrift eines Mitglieds der Universität Duisburg-Essen