Universität Ulm

Institut für Epidemiologie und Medizinische Biometrie

Institutsdirektor: Univ. Prof. Dr. med. Dietrich Rothenbacher, MPH

# Biometrische Aspekte der Fallzahlplanung für Tierversuche in der medizinischen Grundlagenforschung

## - eine empirische Untersuchung -

Dissertation zur Erlangung des Doktorgrades der Medizin der Medizinischen Fakultät der Universität Ulm

Verfasser: Veronika Richter

Geburtsort: Bruchsal

Jahr der Vorlage im Promotionssekretariat: 2017

| | |
|---|---|
| Amtierender Dekan: | Prof. Dr. Thomas Wirth |
| 1. Berichterstatter: | Prof. Dr. Rainer Muche |
| 2. Berichterstatter: | Prof. Dr. med. Dr. med. h. c. Peter Radermacher |
| Tag der Promotion: | 14.06.2018 |

# Inhaltsverzeichnis

**Abkürzungsverzeichnis ..... III**

**1 Einleitung ..... 1**

1.1 Tierversuche in der medizinischen Forschung ..... 1
1.2 Ethische Aspekte ..... 4
1.3 Gesetzeslage für Tierversuchsvorhaben ..... 6
1.4 Das 3R-Prinzip ..... 8
1.5 Fragestellung ..... 9

**2 Material und Methoden ..... 10**

2.1 Tierversuchsgutachten ..... 10
2.2 Der statistische Test ..... 11
2.3 Statistische Testverfahren ..... 18
2.4 Biometrische Fallzahlplanung ..... 25
2.5 Statistikprogramm nQuery Advisor® ..... 26
2.6 Berechnungen ..... 31
2.7 Nicht berücksichtigte Gutachten ..... 35

**3 Ergebnisse ..... 37**

3.1 Vergleich der Fallzahlen – sortiert nach Ausgangskonstellation und betrachtet für die Summe der Jahre 2010 - 2014 ..... 37
3.2 Vergleich der Fallzahlen – sortiert nach Ausgangskonstellation und Jahr ..... 42

**4 Diskussion ..... 47**

4.1 Voraussetzungen ..... 47
4.2 Die Rolle der statistischen Fallzahlplanung ..... 48
4.3 Die Möglichkeiten der statistischen Fallzahlplanung ..... 49
4.4 Das Versuchsdesign und die Verminderung der Tierfallzahl ..... 52
4.5 Die Aussagekraft kleiner Studien ..... 53
4.6 Das 3R-Prinzip und die Verminderung der Tierfallzahl ..... 53
4.7 Ethische Aspekte und die Verminderung der Tierfallzahl ..... 54
4.8 Fazit ..... 55

**5 Zusammenfassung ..... 57**

**6 Literaturverzeichnis ..... 59**

**7 Abbildungs- und Tabellenverzeichnis ..... 65**

**8 Anhang ..... 69**

8.1 Vergleich der Fallzahlen – sortiert nach Ausgangskonstellation und Jahr ..... 69

**Lebenslauf ..... 75**

## Abkürzungsverzeichnis

| | |
|---|---|
| BMEL | Bundesministerium für Ernährung und Landwirtschaft |
| DFG | Deutsche Forschungsgemeinschaft |
| EMA | European Medicines Agency |
| | (dt.: Europäische Arzneimittel-Agentur) |
| RP | Regierungspräsidium |
| TierSchG | Tierschutzgesetz |
| TierSchVersV | Tierschutz-Versuchstierverordnung |

# 1 Einleitung

Dieses Kapitel gibt zunächst einen allgemeinen Überblick über das Tierversuchswesen in der medizinischen Forschung, die zu bedenkenden ethischen Aspekte, sowie die derzeit geltenden Gesetze und Vorschriften. Zudem wird das 3R-Prinzip, die Basis für die Verankerung der biometrischen Fallzahlplanung in der Planung und Durchführung von Tierversuchen, erläutert. Zuletzt wird die dieser Arbeit zugrunde liegende Fragestellung vorgestellt.

## 1.1 Tierversuche in der medizinischen Forschung

Tierversuche finden ihre Begründung im Rahmen der Entwicklung der Medizin von erfahrungsbasierten Methoden hin zu evidenzbasierter, auf Versuchs- und Studienergebnissen beruhender Medizin. Sie sind heute ein wichtiges Instrument der medizinischen Forschung. Aber was versteht man heute eigentlich unter einem Tierversuch? Tierversuche sind laut TierSchG definiert, als „Eingriffe oder Behandlungen an Tieren, wenn diese mit Schmerzen, Leiden oder Schäden für die Tiere verbunden sein können“. Dies gilt für jedes Verfahren, bei dem Tiere einer Belastung ausgesetzt werden, die „dem eines Kanüleneinsatzes gemäß guter tierärztlicher Praxis gleichkommt oder darüber hinaus geht“ (Artikel 3, 2010/63/EU) ([20], S.9). Tierversuche dürfen in der Bundesrepublik Deutschland nach § 7a Abs. 1 TierSchG nur für folgende Zwecke eingesetzt werden: „

1. Grundlagenforschung
2. sonstige Forschung mit einem der folgenden Ziele:
    a) Vorbeugung, Erkennung oder Behandlung von Krankheiten, Leiden, Körperschäden oder körperlichen Beschwerden bei Menschen oder Tieren,
    b) Erkennung oder Beeinflussung physiologischer Zustände oder Funktionen bei Menschen oder Tieren,
    c) Förderung des Wohlergehens von Tieren oder Verbesserung der Haltungsbedingungen von landwirtschaftlichen Nutztieren,
3. Schutz der Umwelt im Interesse der Gesundheit oder des Wohlbefindens von Menschen oder Tieren,

4. Entwicklung und Herstellung sowie Prüfung der Qualität, Wirksamkeit oder Unbedenklichkeit von Arzneimitteln, Lebensmitteln, Futtermitteln oder anderen Stoffen oder Produkten mit einem der in Nummer 2 Buchstabe a bis c oder Nummer 3 genannten Ziele,
5. Prüfung von Stoffen oder Produkten auf ihre Wirksamkeit gegen tierische Schädlinge,
6. Forschung im Hinblick auf die Erhaltung der Arten,
7. Aus-, Fort- oder Weiterbildung,
8. gerichtsmedizinische Untersuchungen.“ [[8], § 7a, Abs. 1 TierSchG].

### *1.1.1 Ablauf des Genehmigungsverfahrens für Tierversuche im Überblick*

Um die Genehmigung für die Durchführung eines Tierversuchs zu erhalten, müssen im entsprechenden Antrag die folgenden drei Elemente dargestellt werden:

Vorhabenbezogen: Es muss eine wissenschaftliche Begründung für das Tierversuchsvorhaben vorgelegt werden. Außerdem müssen die Unerlässlichkeit und die ethische Vertretbarkeit dargestellt werden [19].

Personenbezogen: Es muss nachgewiesen werden, dass sowohl der Projektleiter, wie auch dessen Stellvertreter die benötigten fachlichen Kompetenzen besitzen und sie in der Vergangenheit nicht gegen das Tierschutzgesetz verstoßen haben [19].

Anlagenbezogen: Es muss sichergestellt werden, dass sowohl die personellen, wie auch die baulichen Bedingungen zur Haltung und Pflege der Versuchstiere, sowie für die Durchführung des Tierversuchs erfüllt werden [19].

Werden im vorgelegten Antrag die Voraussetzungen dieser drei Elemente erfüllt, so muss die Genehmigungsbehörde die Erlaubnis zur Durchführung eines Tierversuchsvorhabens innerhalb von drei Monaten erteilen [19]. Bis zu dieser Genehmigung muss der Antrag aber noch die folgenden Schritte durchlaufen:

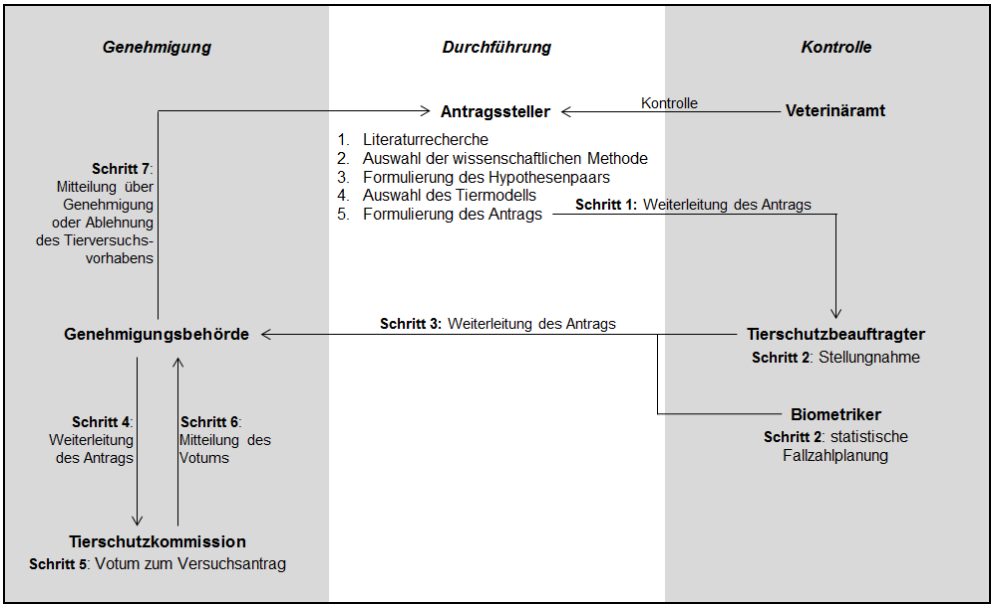


*Abbildung 1: Überblick über den Ablauf des Genehmigungsverfahrens zur Durchführung eines Tierversuchs [19, 39, 46]*

Im ersten Schritt wird der vollständige Antrag vom Antragssteller an den Tierschutzbeauftragten der jeweiligen Einrichtung übermittelt. Dieser nimmt dazu Stellung und leitet ihn, gemeinsam mit der statistischen Fallzahlplanung, an die zuständige Genehmigungsbehörde weiter. Von dort wird der Antrag an die zuständige Tierschutzkommission übermittelt. Diese gibt ein Votum zum entsprechenden Antrag ab, welches für die Genehmigungsbehörde jedoch nicht bindend ist. In einem letzten Schritt teilt die Genehmigungsbehörde dem Antragssteller mit, ob der Antrag auf Durchführung eines Tierversuchsvorhabens genehmigt oder abgelehnt wird [46]. Während der Durchführung des Tierversuchs unterliegen die Wissenschaftler der ständigen internen und externen Kontrolle. Die interne Kontrolle erfolgt durch den Tierschutzbeauftragten der Einrichtung. Die externe Kontrolle hingegen findet durch das zuständige Veterinäramt statt [19].

### *1.1.2 Entwicklung der Tierversuchszahlen zwischen 2010 und 2014*

„Die Verwendung von Wirbeltieren in Tierversuchen oder zu bestimmten anderen wissenschaftlichen Zwecken wird jährlich statistisch erfasst. Rechtliche Grundlage für die Erfassung ist die Versuchstiermeldeverordnung“ [[13], S. 58]. Diese Erfassungspflicht beruht auf der Richtlinie 2010/63/EU Art. 54 Abs. 2, die die Mitgliedsstaaten zur jährlichen Erfassung statistischer Daten über die Verwendung von Tieren und deren Veröffentlichung verpflichtet [16]. Aus diesen Daten ist ersichtlich, dass die Zahl der Tiere, die in Tierversuchen eingesetzt wurden, in den Jahren 2010 bis 2014 nahezu konstant geblieben ist [13, 21].

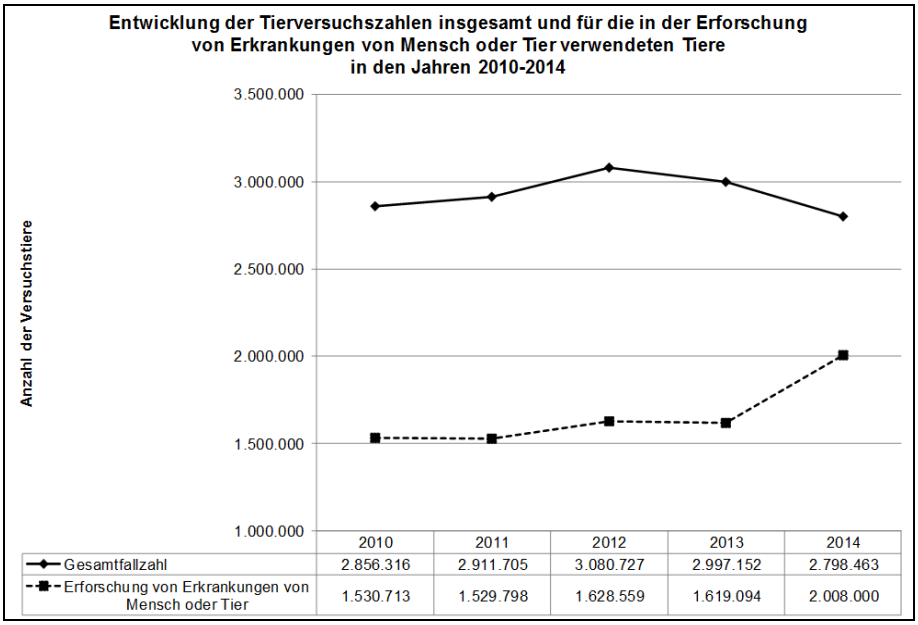


*Abbildung 2: Entwicklung der Tierversuchszahlen in den Jahren 2010-2014 in Deutschland [9-14, 20, 21]*

In den letzten Jahren konnte ein leichter Abfall der Gesamttierfallzahl beobachtet werden. So sank die Zahl im Jahr 2013 um etwa 3% im Vergleich zu 2012 und im Jahr 2014 sogar um ca. 6,6% im Vergleich zum Vorjahr 2013 [20]. Dabei entfielen laut Statistik des BMEL ca. 31,1% der in der Forschung verwendeten Tiere auf die Grundlagenforschung [20]. Betrachtet man die Situation europaweit, so finden sogar 60% der Tiere ihren Einsatz in der Grundlagenforschung [20]. Generell ist zu erwähnen, dass nur 0,35% der in Deutschland in irgendeiner Art und Weise verwendeten Tiere auf die Forschung entfallen [20]. Ein deutlich größerer Anteil, nämlich gut 99% der Tiere, werden für die Herstellung von Lebensmitteln getötet. Hierbei handelt es sich vor allem um Nutztiere wie Rinder, Schweine oder Geflügel, wohingegen in der Forschung bevorzugt Mäuse (68,0%) und Ratten (13,0%) eingesetzt werden [20].

## 1.2 Ethische Aspekte

Damit Projekte ethisch vertretbar sind, müssen diese gut geplant, korrekt durchgeführt und ausgewertet, sowie richtig interpretiert werden [25, 29]. Zudem sollten sie effizient und human, im Sinne einer Durchführbarkeit des Versuchs mit menschlichen Probanden, geplant sein und mit der geringsten möglichen Anzahl an Tieren zum gewünschten Ziel führen [25]. Generell sind „schlechte“ Studien ethisch nicht vertretbar. „Schlechte“ Studien sind solche, bei denen die den Teilnehmern zugemutete Belastung weitaus höher wiegt, als der durch die Studie zu erwartende Erkenntnisgewinn [1]. Wenn allerdings eine Studie

erfolgversprechend scheint, so kann diese auch bei einem „negativen Ergebnis“ ethisch einwandfrei sein. Anders formuliert bedeutet das, dass auch Studien die beispielsweise die Wirksamkeit einer Therapie oder eines Medikaments nicht beweisen konnten – d.h. die Nullhypothese konnte nicht abgelehnt werden – ethisch vertretbar sind, denn solche Studienergebnisse führen in der Zukunft dazu, dass diese Form der Therapie oder dieses Medikament nicht mehr eingesetzt wird, was wiederum einen Vorteil für künftige Patienten darstellt [29].

Andererseits sollten Tierversuche nur dann stattfinden, wenn es keine Alternative gibt, die Ziele stichhaltig sind und bei der Betrachtung des Verhältnisses zwischen der Belastung der Tiere und dem Erkenntnisgewinn dieser überwiegt [8, 16, 25]. Hier bedeutet Stichhaltigkeit, dass das Projekt mit hoher Wahrscheinlichkeit zum gewünschten Ergebnis führt und dieses einen, möglicherweise langfristigen, Einfluss auf das Wohlbefinden von Mensch oder Tier hat [25].

Auch in der „Richtlinie 2010/63/EU des Europäischen Parlamentes und des Rates zum Schutz der für wissenschaftliche Zwecke verwendeten Tiere“ [16] werden ethische Aspekte aufgegriffen. So sollten Tiere „stets als fühlende Wesen behandelt […] und ihre Verwendung […] sollte auf Bereiche beschränkt werden, die […] einen Nutzen für die Gesundheit von Mensch und Tier oder für die Umwelt nach sich ziehen können“ [[16], Einleitung, Abs. 12]. In Anlehnung an das 3R-Prinzip verlangt die europäische Union, dass der Einsatz von Tieren nur dann erfolgen soll, wenn es keine Alternative gibt, die ohne einen Tierversuch auskommt [16]. Zudem soll die Methode gewählt werden, die wahrscheinlich die besten Ergebnisse liefert und so wenig Schmerz, Angst oder Leiden wie möglich verursacht. Ebenso verlangt die Europäische Union, dass so wenige Tiere wie möglich verwendet werden [16]. An diesem Punkt kommt die biometrische Fallzahlplanung ins Spiel.

Auch das Tierschutzgesetz der Bundesrepublik Deutschland beinhaltet einen Absatz zur ethischen Problematik von Tierversuchen. Hier heißt es, dass „Versuche an Wirbeltieren oder Kopffüßern […] nur durchgeführt werden [dürfen], wenn die zu erwartenden Schmerzen, Leiden oder Schäden der Tiere im Hinblick auf den Versuchszweck ethisch vertretbar sind“ [[8], § 7a Abs. 2 Satz 1 Nr. 3 TierSchG]. Zudem beinhaltet dieser Paragraph des Tierschutzgesetzes die Maßgabe, dass immer der aktuelle Stand der Forschung für ein Projekt heran

gezogen werden muss. Zudem muss ausgeschlossen sein, dass das Ziel nicht durch andere Methoden zu erreichen ist und dass den Tieren Schmerzen und Qualen nur im nötigsten Maß zugefügt werden [8].

Ebenso erfordert der Antrag auf Genehmigung eines Tierversuchsvorhabens des Regierungspräsidiums Tübingen, in dessen Zuständigkeit die Universität Ulm fällt, eine wissenschaftlich begründete Darstellung der „ethischen Vertretbarkeit des Versuchs“ [[47], S. 10, Punkt 2.5].

### 1.3 Gesetzeslage für Tierversuchsvorhaben

Die rechtlichen Grundlagen, auf denen die biometrische Fallzahlplanung für die medizinische Forschung beruht, sind in diversen Gesetzen und Vorschriften verankert, die im Folgenden vorgestellt und erläutert werden.

Die primäre Richtlinie, auf der die gesamten rechtlichen Grundlagen in Deutschland beruhen, ist die „Richtlinie 2010/63/EU des Europäischen Parlaments und des Rates vom 22. September 2010 zum Schutz der für wissenschaftliche Zwecke verwendeten Tiere“ [16]. Diese Richtlinie dient zum einen dazu, einen funktionierenden Binnenmarkt zu garantieren und das „Wohlergehen von Tieren“ [[16], Einleitung, Abs. 2] als Wert der Europäischen Union zu unterstreichen. Zum anderen hat „die Gemeinschaft die Bedeutung des Schutzes und des Wohlergehens von Tieren, die für wissenschaftliche Zwecke verwendet werden, auf internationaler Ebene anerkannt“ [[16], Einleitung, Abs. 3]. Des Weiteren wird in dieser Richtlinie darauf verwiesen, dass innerhalb der Europäischen Union „die Pflege und Verwendung lebender Tiere für wissenschaftliche Zwecke […] von den international anerkannten Prinzipien der Vermeidung, Verminderung und Verbesserung bestimmt [wird]“ [[16], Einleitung, Abs. 11]. Dies wird weiter vertieft, in dem die Richtlinie die Mitgliedsstaaten dazu verpflichtet, so wenige Tiere wie nötig zu verwenden, ohne jedoch das Ziel der einzelnen Projekte zu gefährden. Deshalb muss jedes Projekt „im Hinblick auf die Erfüllung der Anforderung der Vermeidung, Verminderung und Verbesserung“ [[16], Artikel 38 Abs. 2b] beurteilt werden.

In der Bundesrepublik Deutschland werden Tierversuchsvorhaben mittels des Tierschutzgesetzes und der Verordnung zum Schutz von zu Versuchszwecken oder anderen wissenschaftlichen Zwecken verwendeten Tieren (Tierschutz-Versuchstierverordnung) geregelt.

Das Tierschutzgesetz in seiner Fassung vom 28.07.2014 erlaubt die Durchführung von Tierversuchen, wenn sie zur Grundlagenforschung unerlässlich sind [8]. Es dient aber auch dazu, das Leben und Wohlempfinden des Tieres als Mitgeschöpf zu schützen, weshalb „niemand […] einem Tier ohne vernünftigen Grund Schmerzen, Leiden oder Schäden zufügen [darf]“ [[8], §1 TierSchG]. Es beinhaltet neben Regelungen zur Tierhaltung, zur Tötung von Tieren, zu Eingriffen an Tieren, zu Tierschutzbeauftragten, zur Zucht und Haltung, sowie zu Straf- und Bußgeldern, auch Bestimmungen zu Tierversuchen. Dieser Abschnitt dient „dem Schutz von Tieren, die zur Verwendung in Tierversuchen bestimmt sind oder deren Gewebe oder Organe dazu bestimmt sind, zu wissenschaftlichen Zwecken verwendet zu werden“ [[8], § 7 Abs. 1 Satz 2 Nr. 1 TierSchG]. Dazu regelt dieser Abschnitt, dass Schmerzen, Leiden und Schäden im Rahmen von Tierversuchen, sowie die Zahl der verwendeten Tiere auf das Minimalste zu senken sind [8].

Die Tierschutzversuchstier-Verordnung in ihrer Fassung vom 31.08.2015 regelt über das Tierschutzgesetz hinaus die Haltung von Versuchstieren, die Durchführung, Genehmigung und Anzeige von Tierversuchen, Ordnungswidrigkeiten, sowie Übergangs- und Schlussbestimmungen [7]. Genauer regelt die Tierschutzversuchstier-Verordnung, dass für den Antrag auf Genehmigung in den einzureichenden Unterlagen „[…] die Anzahl der für das Versuchsvorhaben vorgesehenen Tiere einschließlich deren Berechnung“ [[7], § 31 Abs. 1 Satz 2 Nr. 1c TierSchVersV] enthalten sein muss.

Zur Handhabung von Tierversuchen innerhalb Baden-Württembergs sei an dieser Stelle auf die Stellungnahme des Ministeriums für Ländlichen Raum und Verbraucherschutz vom 05.11.2014 verwiesen. Hier wird erneut auf die Wichtigkeit des 3R-Prinzips als „wichtiges Grundprinzip“ innerhalb der Richtlinie der Europäischen Union und des Tierschutzgesetzes [[18], S. 3, zu 4.] verwiesen.

Die Antragsstellung für die Durchführung eines Tierversuches an der Universität Ulm erfolgt über den Tierschutzbeauftragten der Universität an das Regierungspräsidium Tübingen. Dieser Antrag in seiner Form von Juli 2016 enthält, neben Angaben zum Antragssteller und seiner Qualifikation für die Durchführung von Tierversuchen, auch Angaben zum Versuchsvorhaben. Diese beinhalten außer der Bezeichnung des Versuchsvorhabens, den gültigen Rechtsgrundlagen, den Angaben zum wissenschaftlichen Hintergrund, der

Beschreibung der Eingriffe und der praktischen Durchführung, der ethischen Vertretbarkeit des Projekts und des geplanten Ablaufs nach Versuchsende, auch Angaben zu „Art und Anzahl der vorgesehenen Tiere (§ 31 Abs. 1 Satz 2 Nr. 1c TierSchVersV)“ [[47], S. 7, Punkt 2.3.8]. Genauer fordert der Antrag die „vorgesehene Anzahl und Begründung für die Anzahl der Tiere einschließlich Angaben zur biometrischen Planung“ [[47], S.8, Punkt 2.3.8.3]. Im Rahmen des Antrags werden auch Angaben zum Versuchstyp gefordert. So muss angegeben werden, ob es sich bei dem geplanten Versuch um einen Pilotversuch, einen hypothesengenerierenden oder um einen hypothesentestenden Versuch handelt. Ist letzteres der Fall, so muss zusätzlich das Formblatt „Angaben zur biometrischen Planung“ ausgefüllt werden [[47], S. 8, Punkt 2.3.8.4]. Dieses Formblatt erfordert Angaben zur primären Zielgröße, zur Verteilungsform der Zielgröße, zu Planungsgrößen, statistischem Auswertungsverfahren, Annahmen zu α- und β-Fehler, verwendeten Korrekturen und der Richtung der Testung [48].

Der Antrag auf Genehmigung eines Tierversuchsvorhabens des RP Tübingen, sowie das Formular zur biometrischen Fallzahlplanung stehen auf den Seiten des Tierforschungszentrums Ulm unter dem Punkt „Antragsformulare als Download“ [54] und auf den Seiten des RP Tübingen zur Verfügung.

## 1.4 Das 3R-Prinzip

Wie bereits erwähnt stellt das 3R-Prinzip nach Russell und Burch aus dem Jahre 1959 das zentrale Element zur Begründung der biometrischen Fallzahlplanung für Tierversuchsvorhaben dar [25]. Entstanden ist das 3R-Prinzip aus der Frage heraus, wie Grausamkeit und Rücksichtslosigkeit gegenüber den Tieren vermindert oder beseitigt werden kann [50]. Aus diesem Grund ist es definiert als *Replacement* (Austausch, Ersetzung), *Reduction* (Verminderung) und *Refinement* (Verfeinerung) [2, 25, 50, 51]. Dabei meint der Begriff *Replacement* den Austausch von Tieren gegen andere Methoden oder Materialien, wie beispielsweise Zellkulturen oder Computersimulationen, wann immer dies eine Möglichkeit darstellt [24, 25, 50]. Unter *Reduction* versteht man die Reduzierung der Tierversuchszahl auf das Minimum, das nötig ist, um das gewünschte Ziel mit ausreichender Sicherheit zu erzielen [24, 25, 44, 50]. Dieses Prinzip ist unter sämtlichen möglichen Methoden diejenige, die sofort einen Nutzen im Hinblick auf die Effizienz in Bezug auf den Einsatz von Tieren bringt [2]. Durch diesen Punkt

findet die statistische Fallzahlplanung ihren Platz im Rahmen von Tierversuchsvorhaben [25]. *Refinement* beschreibt den Prozess, grausame und unerträgliche, aber im Rahmen des Projekts unvermeidbare Schmerzen, Qualen oder Prozeduren in Häufigkeit und Schwere auf ein Mindestmaß zu reduzieren [24, 50]. Enthalten im Begriff Refinement sind beispielsweise eine Betäubung und Schmerzlinderung bei chirurgischen Eingriffen oder eine dem jeweiligen Tier entsprechende Unterbringung und Versorgung durch ausgebildetes Personal. Als grausam, unerträglich oder sogar barbarisch ist beispielsweise ein auf stundenlange Qualen hin folgender Tod zu bezeichnen. Dies sollte wenn möglich vermieden werden [24, 25]. Ein alleiniges *Refinement* gilt als nicht ausreichend, weshalb als Ziel besser *Reduction* und wann immer möglich *Replacement* verfolgt werden sollte [2]. Natürlich überlappen sich die drei Bestandteile, aber die vorgeschlagene Dreiteilung ist sicherlich sinnvoll, um dieses Problem zu strukturieren und die Projektleiter zum Überdenken ihrer Projektplanung anzuregen [2, 50]. Trotzdem erscheint es prinzipiell sinnvoll die beiden Prinzipien *Reduction* und *Refinement* zu kombinieren, v.a. für den Fall, dass ein *Replacement* nicht möglich ist [2].

## 1.5 Fragestellung

Unter Berücksichtigung der Grundsätze des 3R-Prinzips, den derzeit geltenden rechtlichen Grundlagen und der Aussagekraft des Ergebnisses eines statistischen Tests, soll diese Arbeit die Frage klären, inwieweit sich die Fallzahl eines Experiments in der medizinischen Grundlagenforschung reduzieren lässt, ohne dass das Ergebnis an Aussagekraft, im Sinne eines Powerverlustes oder eines zu großzügigen Signifikanzniveaus, verliert. Hierzu werden die verschiedenen Szenarien der biometrischen Fallzahlplanung mit Hilfe eines Statistikprogramms berechnet und anschließend ausgewertet. Für die Auswertung bedarf es einer kritischen Auseinandersetzung mit den derzeit geltenden Standards für Auswertung und Aussagekraft statistischer Tests. Ebenso muss sich mit der Frage beschäftigt werden, ob solch hohe Anforderungen bereits in der Grundlagenforschung von Nöten sind, oder ob es nicht ausreicht, diese Standards erst in einer späteren Phase des Studiendesigns anzuwenden. Dies gilt es vor allem unter dem Gesichtspunkt zu betrachten, dass die ersten Versuchsreihen meist zur Klärung der Frage, ob sich weitere Forschung lohnt, dienen.

# 2 Material und Methoden

Dieses Kapitel stellt zum einen die dieser Arbeit zugrunde liegenden Tierversuchsgutachten des Instituts für Epidemiologie und Medizinische Biometrie der Universität Ulm, die unterschiedlichen Formen der Auswertung der durchgeführten Berechnungen, die möglichen Fehlerquellen, die nicht berücksichtigten Gutachten und das Statistikprogramm *nQuery Advisor*® vor. Zum anderen geht dieses Kapitel auf die verschiedenen Komponenten eines statistischen Tests und die gängigsten statistischen Testverfahren ein.

## 2.1 Tierversuchsgutachten

Der Auswertung liegen insgesamt 164 Tierversuchsgutachten mit einem Umfang von insgesamt 76.620 berechneten Tieren im Zeitraum vom 01.01.2010 bis zum 31.12.2014 des Instituts für Epidemiologie und Medizinische Biometrie der Universität Ulm zu Grunde. Diese Zahlen beruhen auf der Gesamtzahl aller in diesem Zeitraum bearbeiteten Tierversuchsgutachten. Enthalten in der Gesamtzahl der Gutachten sind 27 Pilotstudien, für die keine Berechnung stattfinden konnte und die nicht in der Endauswertung berücksichtigt werden. Zudem sind 25 weitere Gutachten inbegriffen, die aus verschiedenen Gründen (s. 2.7. Nicht berücksichtigte Gutachten) nicht weiter verfolgt wurden. Somit werden in der abschließenden Auswertung 112 Tierversuchsgutachten, die im Ulmer Institut für Epidemiologie und Medizinische Biometrie erstellt wurden, mit einer Gesamtzahl von 59.331 berechneten Tieren berücksichtigt.

Die Gesamtsumme der genannten 76.620 Tiere verteilt sich wie folgt:

*Tabelle 1: Gesamtsumme der Tiere, aufgeschlüsselt nach Jahr und Gattung*

| | **2010** | **2011** | **2012** | **2013** | **2014** | **gesamt** |
|---|---|---|---|---|---|---|
| Mäuse | 17.842 | 11.481 | 14.388 | 4.935 | 23.620 | **72.266** |
| Ratten | 1.143 | 46 | 0 | 0 | 880 | **2.069** |
| Kaninchen | 91 | 0 | 104 | 14 | 0 | **209** |
| Schafe | 98 | 77 | 36 | 51 | 75 | **337** |
| Schweine | 68 | 0 | 126 | 9 | 60 | **263** |
| Ferkel | 0 | 34 | 0 | 37 | 0 | **71** |
| Singvögel | 0 | 600 | 0 | 0 | 675 | **1.275** |
| Siebenschläfer | 0 | 80 | 0 | 0 | 0 | **80** |
| Krallenfrösche | 0 | 0 | 0 | 50 | 0 | **50** |
| **gesamt** | **19.242** | **12.318** | **14.654** | **5.096** | **25.310** | **76.620** |

Aus Tabelle 1 wird ersichtlich, dass vor allem Mäuse (94,3%) als Versuchstiere an der Universität Ulm eingesetzt werden. Ratten finden sich mit ca. 2.000 Individuen und einem prozentualen Anteil von 2,7% auf Platz 2. Siebenschläfer, Krallenfrösche und Ferkel stellen hingegen zahlenmäßig eher eine Ausnahme dar.

Die in der Endberechnung berücksichtigten 59.331 Tiere gliedern sich folgendermaßen:

*Tabelle 2: Anzahl der in der Endberechnung berücksichtigten Tiere, aufgeschlüsselt nach Jahr und Gattung*

| | **2010** | **2011** | **2012** | **2013** | **2014** | **gesamt** |
|---|---|---|---|---|---|---|
| Mäuse | 8.234 | 10.135 | 14.312 | 4.600 | 19.063 | 57.331 |
| Ratten | 1.143 | 46 | 0 | 0 | 352 | 1.541 |
| Kaninchen | 28 | 0 | 28 | 0 | 0 | 56 |
| Schafe | 98 | 77 | 28 | 0 | 42 | 245 |
| Schweine | 68 | 0 | 34 | 9 | 0 | 111 |
| Ferkel | 0 | 34 | 0 | 0 | 0 | 34 |
| **gesamt** | **9.571** | **11.292** | **14.402** | **4.609** | **19.457** | **59.331** |

Auch anhand dieser Tabelle, die die in der Endberechnung berücksichtigten Tiere nach Jahr und Gattung auflistet, lässt sich bestätigen, dass überwiegend Mäuse (96,6%), mit großem Abstand gefolgt von Ratten (2,6%), für Tierversuchsvorhaben im Rahmen der medizinischen Grundlagenforschung an der Universität Ulm eingesetzt werden. Singvögel, Krallenfrösche und Siebenschläfer hingegen fanden lediglich Verwendung im Rahmen der ökologischen Forschung, weshalb diese Tiergattungen in der Endberechnung keine Berücksichtigung fanden. Dies begründet sich unter anderem in der Tatsache, dass im Rahmen der ökologischen Forschung keine Tötung der Tiere stattfindet, sondern diese in der freien Natur gefangen und nach Abschluss der Versuche wieder freigelassen werden.

## 2.2 Der statistische Test

Ein statistischer Test dient primär dazu, eine aufgestellte Hypothese zu beweisen [44]. Diese Überprüfung erfolgt auf zwei Arten: Zuerst muss ein theoretischer Hintergrund erarbeitet werden, um die Hypothese sachlich argumentiert zu untermauern. Des Weiteren ist es von Nöten, die aufgestellte Hypothese mittels Datenerhebung und statistischer Analyse zu bestätigen [56]. Um einen statistischen Test zur Anwendung bringen zu können, müssen zunächst Null- und

Alternativhypothese formuliert werden. Dies muss zwingend vor der Durchführung des Testes geschehen [56].

Dabei gilt:

$H_0$ = Nullhypothese

$H_1$ = Alternativhypothese

Diese beiden Hypothesen verhalten sich komplementär zueinander, d.h. sie schließen sich gegenseitig aus bzw. ergänzen sich [56]. Im Rahmen des statistischen Tests soll die Nullhypothese richtigerweise abgelehnt und somit die Alternativhypothese indirekt bestätigt werden. Die Hypothesen werden dabei so formuliert, dass alle denkbaren Möglichkeiten abgedeckt und dadurch Aussagen über die Grundgesamtheit getroffen werden können. Im Regelfall beinhaltet die Nullhypothese die Gleichheit (Ausnahme: Äquivalenztests) und die Alternativhypothese den Unterschied [4, 30]. Prinzipiell lässt sich mit Hilfe eines statistischen Tests jede zweiseitig formulierte Alternativhypothese bestätigen, wenn nur der Stichprobenumfang groß genug gewählt wird [15, 56].

Für die Berechnung der für ein signifikantes Testergebnis jeweils benötigten Tierzahlen sind drei statistische Variablen von entscheidender Bedeutung: Das Signifikanzniveau α, die statistische Power 1-β, sowie das Hypothesenpaar, das je nach Fragestellung ein- oder zweiseitig sein kann.

### *2.2.1 Fehlerarten*

Für die Erklärung der Begrifflichkeiten sollen kurz die Fehlerarten des statistischen Tests erläutert werden:

*Tabelle 3: Wahrheitstafel [56]*

| | | In Wahrheit gilt: | |
|---|---|---|---|
| | | $H_0$ | $H_1$ |
| Entscheidung für: | $H_1$ | Fehler 1. Art (α) | richtig (1-β) |
| | $H_0$ | richtig (1-α) | Fehler 2. Art (β) |

Entscheidet man sich für die Alternativhypothese, obwohl die Nullhypothese richtig ist, so liegt ein Fehler 1. Art (α-Fehler) vor [4]. Der Test wird falsch signifikant bzw.

falsch-positiv und behauptet einen Unterschied, der in Wahrheit nicht existiert [30]. Dieser Fehler kann nicht vermieden, aber kontrolliert werden, da er nur bei Gültigkeit der Nullhypothese auftritt, welche jedoch eindeutig formuliert ist [56].

Trifft man die Entscheidung zu Gunsten der Nullhypothese, obwohl die Alternativhypothese Gültigkeit hat, so liegt ein Fehler 2. Art (β-Fehler) vor [4]. In diesem Fall ist der Test fälschlicherweise nicht signifikant, also falsch-negativ, was wiederum bedeutet, dass der Test einen in Wahrheit bestehenden Unterschied nicht erkannt hat [17, 30]. Ein Fehler 2. Art ist, im Gegensatz zu einem Fehler 1. Art, kaum abschätzbar, da im Gegensatz zur Nullhypothese die Alternativhypothese nicht eindeutig vorgegeben ist [56].

Da der β-Fehler durch die Wahl des α-Fehlers beeinflusst wird, gilt: Die Wahl eines strengeren Signifikanzniveaus α bedeutet auf der einen Seite, dass seltener die gültige Nullhypothese abgelehnt wird. Auf der anderen Seite bedeutet dies aber auch, dass mit einer größeren Wahrscheinlichkeit die Nullhypothese auch dann für richtig angenommen wird, wenn in Wahrheit die Alternativhypothese Gültigkeit hat [4, 56].

### *2.2.2 Signifikanzniveau α*

Als Signifikanzniveau α wird die Grenzwahrscheinlichkeit bezeichnet, ab der das gefundene Ergebnis nicht mehr als zufällig entstanden betrachtet wird [17, 30]. Das Signifikanzniveau α eines statistischen Tests ist prinzipiell frei wählbar, wird aber in der Regel auf 5% gesetzt, um „eine bessere Vergleichbarkeit statistisch abgesicherter Entscheidungen zu ermöglichen“ [[56]; Seite 182]. Generell sollte das Signifikanzniveau vor der Durchführung des Tests festgelegt werden [56]. Ein Signifikanzniveau α von 5% bedeutet, dass die Irrtumswahrscheinlichkeit des Tests, oder die Wahrscheinlichkeit für einen Fehler 1. Art, bei 5% liegt [25]. Daraus ergibt sich die Konsequenz, dass man sich mit einer Wahrscheinlichkeit von 5% für die Alternativhypothese $H_1$ entscheidet, obwohl kein entsprechender Effekt vorliegt und die Nullhypothese Gültigkeit hat [17]. Der Fehler 1. Art wird somit umso kleiner, je niedriger der Wert für α gewählt wird. Dies bedeutet, dass je kleiner das gewählte Signifikanzniveau, desto genauer das Ergebnis des Tests. Zudem wird das Signifikanzniveau herangezogen, um zu entscheiden, ob das Ergebnis des Tests „statistisch signifikant“ ist. Dazu wird der angenommene Wert für α mit dem p-Wert des Tests verglichen. Der p-Wert kann mittels einer

geeigneten Software anhand der Testergebnisse berechnet werden und ist definiert als „Wahrscheinlichkeit, mit der die tatsächlich gewonnenen Daten – oder noch extremere, noch unwahrscheinlichere Daten – rein zufällig unter der Nullhypothese entstanden sein können“ [[30], S. 266]. Ist der berechnete p-Wert also sehr klein, so ist es unwahrscheinlich, dass die Nullhypothese gilt. Sie wird verworfen und die Alternativhypothese angenommen [30]. Für ein gewähltes Signifikanzniveau α von 5% bedeutet das, wenn der p-Wert kleiner oder gleich dem gewählten Signifikanzniveau α ist, dass die Nullhypothese mit einer Irrtumswahrscheinlichkeit von lediglich 5% abgelehnt und die Alternativhypothese angenommen wird [56].

Zu beachten ist jedoch, dass der p-Wert nur eine Aussage darüber trifft, ob ein statistisch signifikanter Unterschied besteht; er macht aber keine Angaben über dessen Größe [56].

Sonderfall: Multiples Testen

In der Praxis besteht gerne die Tendenz, einen Test nach dem anderen durchzuführen, um dadurch zumindest ein signifikantes Ergebnis zu erhalten [56]. Dabei gilt zu berücksichtigen, dass im Rahmen des sogenannten „multiplen Testens“ der α-Fehler sehr stark ansteigt und somit auch die Wahrscheinlichkeit, dass das Ergebnis zufällig entstanden ist [56]. Das bedeutet gleichzeitig, dass ein deutlich größeres Risiko besteht die Nullhypothese fälschlicherweise abzulehnen. Um dieser Gefahr entgegen zu wirken geht man in der Statistik folgendermaßen vor: Generell gilt, dass die Wahrscheinlichkeit sich unter Gültigkeit der Nullhypothese bei einem einzeln durchgeführten Test richtig zu entscheiden bei (1-α) liegt [56]. Des Weiteren gilt, dass diese Wahrscheinlichkeit bei beispielsweise 3 unabhängig durchgeführten Tests nur noch bei $(1-\alpha)^3$ liegt. Bei einem gewählten Signifikanzniveau α von 5% entspricht diese Wahrscheinlichkeit etwa 85%, womit der Fehler 1. Art nahezu 15% beträgt. Aus diesem Grund wird beim multiplen Testen die sogenannte Bonferroni-Korrektur angewendet, wonach ein einzelnes Testergebnis erst dann als signifikant betrachtet wird, wenn der ermittelte p-Wert kleiner ist als α/k [56]. Mathematisch kann gezeigt werden, dass der Fehler 1. Art bei einer Anzahl von k durchgeführten Tests berechnet werden kann als k*α [56]. Aus diesem Grund muss umgekehrt auch das „lokale“ Signifikanzniveaus α für jeden Test niedriger gewählt werden, um insgesamt den

Fehler 1. Art (globaler Fehler) auf α zu beschränken. Im oben beschrieben Fall kann eine gewünschte Wahrscheinlichkeit eines Fehlers 1. Art von 5% erzielt werden durch die Wahl eines Signifikanzniveaus α von 1,67%. Dieses Signifikanzniveau α berechnet sich als: Fehler 1. Art = 5%, k = 3, woraus folgt α = 5% / 3 = 1,67%. An dieser Stelle sei bereits auf die dieser Doktorarbeit zugrunde liegenden Tierversuchsgutachten verwiesen. Bei Ausgangskonstellation 4 (siehe 2.6.1 Ausgangskonstellationen) wurde im Rahmen der biometrischen Fallzahlplanung das Signifikanzniveau mit α = 1,67% angenommen. Dies geschah mit dem Ziel, dass, bei geplantem multiplem Testen im Rahmen der Versuchsdurchführung, der Fehler 1. Art nicht mehr als 5% beträgt. Zu beachten gilt allerdings, dass bei Anwendung der Bonferroni-Korrektur der β-Fehler stark ansteigt [56].

### *2.2.3 Power 1-β*

Die Power (auch: Güte, Teststärke, Trennschärfe oder Macht) 1-β eines statistischen Tests bezeichnet dessen Aussagekraft. Sie gibt die Wahrscheinlichkeit an, mit der ein statistischer Test einen tatsächlich vorhandenen Unterschied und damit eine gültige Alternativhypothese zu einem gegebenen Signifikanzniveau α, erkennt [4, 17, 25, 30, 56]. Die Power leitet sich aus dem Fehler 2. Art (β-Fehler) ab, der die Wahrscheinlichkeit angibt, die Alternativhypothese fälschlicherweise abzulehnen und damit die Nullhypothese anzunehmen. Dies würde bedeuten, dass der Test fälschlicherweise nicht signifikant wird. Das Risiko für einen Fehler 2. Art ist sehr groß und kann nicht kontrolliert werden. Die Power wird aber umso größer, je größer die Fallzahl, der tatsächliche, wirkliche Unterschied zwischen den Gruppen oder der Zielgröße und das Signifikanzniveau werden [15, 44, 56].

Da es in der Forschung durchaus schlimmere Folgen hat, eine falsch positive Behauptung (Fehler 1. Art) zu machen als eine falsch Negative (Fehler 2. Art), gibt es die Vereinbarung das Signifikanzniveau α auf 5% und die Power 1-β auf 80%, d.h. β auf 20%, zu setzen. Daraus ergibt sich, dass ein Fehler 2. Art vierfach wahrscheinlicher ist, als ein Fehler 1. Art. Dies ist aufgrund der Schwere eventueller Folgen durchaus begründbar [15]. Deshalb sollte ein statistischer Test so konstruiert werden, dass das Signifikanzniveau unabhängig vom

Stichprobenumfang konstant bleibt und sich die Power 1-β je nach Stichprobengröße verändert.

### *2.2.4 Hypothesenpaare*

Bei der dritten genannten Variablen handelt es sich um das Hypothesenpaar. Generell wird zwischen einem ein- und einem zweiseitigen Hypothesenpaar unterschieden. Die Festlegung, ob die Fragestellung ein- oder zweiseitig formuliert wird, muss vor Testbeginn erfolgen. Besteht Unsicherheit bezüglich der Voraussetzungen für eine einseitige Testung, so ist es sinnvoller eine zweiseitige Testung durchzuführen [56]. Allerdings benötigt diese eine geringfügig größere Fallzahl [3, 25]. Egal, ob es sich bei der Fragestellung um ein einseitiges oder zweiseitiges Hypothesenpaar handelt, enthält die Nullhypothese stets die Gleichheit (Ausnahme: Äquivalenztests) [30]. Außerdem gilt es zu beachten, dass ein zweiseitiger Test für jedes beliebig gewählte α strenger ist als ein einseitiger Test [15].

<u>einseitiges Hypothesenpaar</u>

Im Falle eines einseitigen oder gerichteten Hypothesenpaares besteht das Ziel des statistischen Testes darin, den Unterschied in einer vor Testbeginn festgelegten Richtung zu entdecken. In diesem Fall beinhaltet die Alternativhypothese nur eine Seite, wohingegen die andere Seite in der Nullhypothese enthalten ist [30]. Für die Anwendung einer einseitigen Testung ist mehr Vorwissen erforderlich, da bereits vor der statistischen Testung die Richtung des möglichen Unterschieds bekannt sein muss. Aus diesem Grund ist für eine einseitige Fragestellung vor Testbeginn eine Tendenzprüfung innerhalb der erhobenen Daten von Nöten. Geht diese Tendenz allerdings in die falsche Richtung, also in Richtung der Nullhypothese, so gilt per Definition, dass der Test nicht signifikant ist und gar nicht erst ausgewertet werden muss. Zudem hat eine einseitige Fragestellung bereits zu Beginn ein höheres Informationsniveau und führt auch schneller zum Ergebnis, allerdings mit der Gefahr an der genannten Tendenzprüfung zu scheitern [30].

Das einseitige Hypothesenpaar stellt sich formal so dar:

$H_0$: $A \leq B$ („Gruppe A ist nicht stärker/besser/schwerer/… als Gruppe B“)

$H_1$: $A > B$ („Gruppe A ist stärker/besser/schwerer/… als Gruppe B“)

Bei einer einseitigen Fragestellung wird die Nullhypothese in der Regel gegensätzlich zur Alternativhypothese formuliert [56].

zweiseitiges Hypothesenpaar

Handelt es sich um ein zweiseitiges oder ungerichtetes Hypothesenpaar, dient der statistische Test dazu, einen Unterschied zu entdecken, unabhängig von dessen Richtung. Dies bedeutet, dass die Nullhypothese $H_0$ eindeutig formuliert ist, wohingegen die Alternativhypothese $H_1$ alle weiteren Hypothesen, mit Ausnahme der Nullhypothese, einschließt [56]. Formal wird eine zweiseitige Fragestellung wie folgt dargestellt:

$H_0$: A = B („die Gruppen unterscheiden sich nicht hinsichtlich der Zielgröße“)

$H_1$: A ≠ B („die Gruppen unterscheiden sich hinsichtlich der Zielgröße“)

Die Bezeichnung zweiseitige Fragestellung rührt von der Tatsache her, dass der Ausdruck der Alternativhypothese $H_1$: A ≠ B beide Seiten, d.h. sowohl A > B, als auch A < B beinhaltet [30].

Einfluss des Hypothesenpaares auf das Signifikanzniveau

Für jeden statistischen Test gibt es eine Menge möglicher Testwerte, die in ihrer Gesamtheit als Datenraum bezeichnet werden. Die durch den Tester getroffene Entscheidung hinsichtlich der Annahme oder Ablehnung der Nullhypothese teilt diesen Datenraum in zwei Bereiche: Den Annahme- und den Ablehnungsbereich [41]. Der Annahmebereich ist definiert als 1-α, der Ablehnungsbereich als α/2 [56].

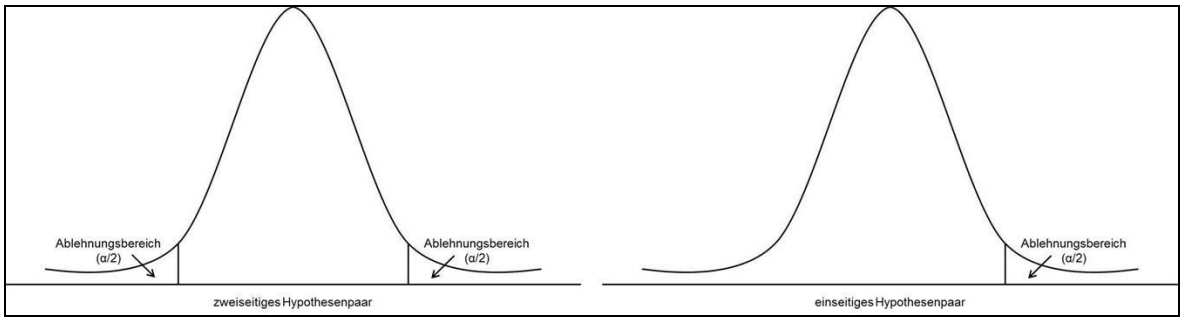


*Abbildung 3: Annahme - und Ablehnungsbereich beim t-Test; nach Vorlage von [41, 56]*

Bei einer zweiseitigen Fragestellung wird, wie bereits bekannt, hinsichtlich eines Unterschieds, oder anders ausgedrückt, hinsichtlich der Gleichheit getestet. In diesem Fall gibt es zwei mögliche Ablehnungsbereiche. Das bedeutet, wenn die Wahrscheinlichkeit der Lage eines Testwertes im Annahmebereich bei 95% liegt,

erhält man zwei Ablehnungsbereiche mit einem Wert von jeweils 2,5% und somit ein Signifikanzniveau von 5% [41, 56].

Im Gegensatz dazu testet man bei einer einseitigen Fragestellung die Richtung des Unterschieds. Das bedeutet, dass eine einseitige Fragestellung nur einen Ablehnungsbereich hat [41]. Geht man von oben genannter Konstellation aus, so erhält man einen Ablehnungsbereich mit 2,5% und einen Annahmebereich von 97,5%. Daraus ergibt sich ein Signifikanzniveau von 2,5% [43].

## 2.3 Statistische Testverfahren

Für die Berechnung der einzelnen Tierversuchsgutachten kommen verschiedene statistische Testverfahren zum Einsatz. Als absolute Voraussetzung für die Anwendung aller im Folgenden genannten Testverfahren gilt, „dass eine zufällige Stichprobe vorliegt, die repräsentativ für die Grundgesamtheit ist“ [[57]; S. 159]. Die in den vorliegenden Gutachten am häufigsten zum Einsatz gekommenen Tests sollen im Folgenden erklärt werden.

### *2.3.1 t-Test*

Der t-Test gehört zu den Lokalisations- oder Lagetests und wird zum Vergleich von Mittelwerten verwendet [58]. Das bedingt, dass es sich bei den zu vergleichenden Lageparametern um stetige, normalverteilte Merkmale handeln muss [25, 31]. Da bei bekannter Verteilung der Variablen nur noch bestimmte Parameter geprüft werden müssen, werden t-Tests auch als parametrische Tests bezeichnet [58].

t-Tests gibt es für eine Stichprobe, für zwei unverbundene oder zwei verbundene Stichproben [58]. Diese Unterscheidung bedingt Unterschiede bei den Voraussetzungen, dem Testverfahren, der Tendenzprüfung und schlussendlich auch der Fallzahlplanung. Generell gilt aber, dass „der t-Test eine stetige, normalverteilte Zielgröße“ voraussetzt [[31], S. 321]. Diese Voraussetzung gilt unabhängig davon, welche Form des t-Test angewendet werden soll.

<u>t-Test für eine Stichprobe</u>

Diese Form des t-Tests dient zum Vergleich des Mittelwerts einer Stichprobe mit einem Modell- oder Sollwert, was voraussetzt, dass die Stichprobenwerte normalverteilt sind [58]. Dieser Test ist allerdings mit Vorsicht anzuwenden. Bei

einer Stichprobegröße von n < 10 sollten, v.a. dann wenn keine gesicherte Normalverteilung vorliegt, andere Lagetests, wie z.B. der Wilcoxon-Test für eine Stichprobe zur Auswertung herangezogen werden [58].

t-Test für zwei unverbundene Stichproben (Paralleldesign)

Der t-Test für zwei unverbundene Stichproben vergleicht die Lageparameter in Form der Mittelwerte mehrerer Grundgesamtheiten miteinander [58]. Für diese Form des t-Tests muss eine unverbundene Versuchsanordnung vorliegen. Das bedeutet, es liegen zwei voneinander unabhängige Stichproben vor. Die Nullhypothese ist definiert als die Gleichheit bezüglich der Merkmale innerhalb der beiden Gruppen - die Alternativhypothese als Unterschied [58]. Der t-Test für unverbundene Stichproben kann für ein- und für zweiseitige Fragestellungen angewendet werden [31].

Soll dieser t-Test zur Anwendung kommen, bedeutet das, dass in beiden Versuchsgruppen n unabhängige Beobachtungen stattfinden, die Zielgröße normalverteilt, mindestens eingipflig und symmetrisch ist und eine Varianzhomogenität vorliegt [31, 58]. Des Weiteren sollten in jeder der beiden Gruppen annähernd gleich viele Beobachtungen gemacht werden. Idealerweise mit einem Stichprobenumfang von mehr als zehn Beobachtungen [31, 58].

Um das Testverfahren zur Anwendung bringen zu können, wird zunächst „für jede Gruppe getrennt die Abweichung der Einzelwerte von ihrem Gruppenmittel bestimmt“ [[31], S. 322], um anschließend die gemeinsame, gepoolte Standardabweichung, die gemeinsame Varianz und die Testgröße zu bilden [31]. Diese kann mit entsprechenden Werten aus der sogenannten t-Tabelle verglichen werden. Ist die ermittelte Testgröße größer oder gleich dem kritischen Wert der t-Tabelle, kann die Nullhypothese abgelehnt werden [31].

Für die Testentscheidung gilt folgendes: Ein großer Unterschied in der beobachteten Differenz der Mittelwerte, eine kleine Varianz innerhalb der Gruppen, eine große Fallzahl, ein hohes Signifikanzniveau, sowie eine einseitige Fragestellung führen generell schneller zu Signifikanz [31].

Wird diese Form des t-Tests auf eine einseitige Fragestellung angewendet, muss eine Tendenzprüfung stattfinden. Das bedeutet, dass untersucht werden muss, ob die Differenz der Mittelwerte mit der Alternativhypothese übereinstimmt. Ist dies

der Fall, wird der t-Test zum einseitigen Signifikanzniveau durchgeführt. Ist dies nicht der Fall, wird das Ergebnis nicht signifikant [31].

Soll für die Auswertung eines Versuchs ein t-Test für zwei Parallelgruppen angewendet werden, müssen das Signifikanzniveau, der Unterschied zwischen den Mittelwerten der beiden Gruppen und die Power vorgegeben werden. Zudem muss die vermutete gemeinsame Standardabweichung der beiden Gruppen angegeben sein. Sind diese Daten vorhanden, kann die benötigte Fallzahl n pro Gruppe berechnet werden [32].

t-Test für zwei verbundene Stichproben (Wertepaare)

Eine verbundene Versuchsanordnung liegt vor, wenn verbundene Stichproben, d.h. Wertepaare verglichen werden sollen (z.B. im Rahmen sogenannter „Vorher-Nachher-Vergleiche“ [[58], S.197]). In diesem Fall dient der t-Test dem Vergleich von Erwartungswerten in Form von Lageparametern der beiden verbundenen Stichproben. Das bedeutet, dass für diese Form des t-Tests n unabhängige Wertepaare vorliegen, die jeweils aus den entsprechenden Grundgesamtheiten mit den zugehörigen Erwartungswerten $\mu_1$ und $\mu_2$ stammen [58]. „Er prüft, ob der Mittelwert der individuellen Differenzen null ist (Nullhypothese) oder nicht (Alternative)“ [[31], S.324]. Auch diese Form des t-Tests kann für ein- und zweiseitige Fragestellungen verwendet werden [31]. Die Anwendung bedingt, dass die Differenzen der Wertepaare eingipflig symmetrisch um einen unbekannten Mittelwert verteilt sind [31].

Bei Anwendung des t-Tests für Wertepaare wird zunächst das Signifikanzniveau gewählt. Es werden die Differenzen der Wertepaare gebildet und anschließend deren Mittelwerte und Standardabweichungen bestimmt. Auch bei dieser Form des t-Tests wird die Prüfgröße ermittelt und mit der t-Tabelle verglichen. Ebenso gilt auch hier, dass die Nullhypothese abgelehnt wird, falls die ermittelte Prüfgröße größer oder gleich dem kritischen Wert der t-Tabelle ist.

Bei Anwendung des t-Tests für ein Wertepaar, also bei einer einseitigen Fragestellung, muss ebenfalls eine Tendenzprüfung stattfinden. Dabei wird untersucht, ob der beobachtete Mittelwert der Differenzen mit der Alternativhypothese übereinstimmt. Ist dies der Fall, wird der Test zum einseitigen Signifikanzniveau durchgeführt. Stimmt der Mittelwert der Differenzen nicht mit der Alternativhypothese überein, dann wird der Test nicht signifikant [31].

### *2.3.2 Welch-Test*

Der Welch-Test (auch: Welch-Korrektur oder t-Test nach Satterthwaite), gehört ebenfalls zu den Lagetests und dient als Alternative zum t-Test für zwei unverbundene Stichproben. Im Vergleich zum t-Test ist für die Anwendung des Welch-Tests das Vorliegen einer Varianzhomogenität keine Bedingung [58]. Zu beachten ist jedoch, dass der Welch-Test bei bekannten Varianzen eine geringere Power hat, als der t-Test für unverbundene Stichproben. Aus diesem Grund sollte in einem solchen Fall der t-Test bevorzugt werden [58]. Außerdem sollte beachtet werden, dass beim Vergleich von zwei Stichproben mit ungleichen Varianzen und damit mit verschiedener Verteilung der Merkmale auch unterschiedliche Erwartungswerte vorliegen und somit eine Vergleichbarkeit nicht sicher gegeben oder sinnvoll ist [58]. Deshalb sollte der Welch-Test nur in begründeten Ausnahmefällen zur Anwendung kommen.

### *2.3.3 Wilcoxon-Test*

Für den Fall, dass die Voraussetzungen für die Anwendung eines t-Tests nicht erfüllt werden, kann ein Rangsummentest, wie der Wilcoxon-Test angewendet werden. Rangsummentests haben schwächere Voraussetzungen als t-Tests. Sie beruhen auf dem Vergleich von in den Originaldaten enthaltenen Rangsummen-informationen und vergleichen zwei Mediane miteinander [58]. Sie können prinzipiell zur Auswertung von quantitativen, Anzahl-, Rang- oder Ratingmerkmalen eingesetzt werden [33].

Der Wilcoxon-Test stellt die nicht parametrische Variante des t-Tests dar und kann unter Umständen auch zur Auswertung ordinal-skalierter Merkmale verwendet werden [5, 58]. Er dient der Auswertung von stetigen, aber nicht normalverteilten Merkmalen mit einer symmetrischen Verteilung [58]. Die Anwendung des Wilcoxon-Tests ist v.a. bei Vorliegen einer schiefen Verteilung und dem Vorhandensein von Ausreißern sinnvoll [33]. Die Anwendung des Wilcoxon-Tests erzeugt für den Fall, dass zwei oder mehrere gleiche Werte auftreten, Bindungen in den Daten, da diesen Werten derselbe Rang (genauer der Mittelwert der beteiligten Ränge) zuzuordnen ist [33, 56]. Sind innerhalb eines Datensatzes 10% oder mehr der Werte an Bindungen beteiligt, so muss eine Bindungskorrektur erfolgen, um einen zu großen Powerverlust zu vermeiden. Ist das Ergebnis des

Wilcoxon-Tests schon vor Anwendung der Bindungskorrektur signifikant, kann diese entfallen und das Ergebnis ist als signifikant zu werten [33].

Wilcoxon-Test für eine Stichprobe

Wie schon der t-Test für eine Stichprobe vergleicht auch der entsprechende Wilcoxon-Test die Gruppe gegen einen Sollwert – nur handelt es sich im Falle eines Wilcoxon-Tests, im Gegensatz zum t-Test, dabei um den Median [58]. Die Nullhypothese ist somit vergleichbar zum t-Test definiert als Gleichheit. Die Alternativhypothese somit als Unterschied [58]. Für die Auswertung werden die Differenzen der Stichprobenwerte zum Sollwert berechnet, Rangzahlen vergeben, Rangsummen gebildet und die Prüfgröße ermittelt. Ist diese ermittelte Prüfgröße kleiner oder gleich dem kritischen Wert, so wird die Nullhypothese abgelehnt [58].

Wilcoxon-Test für zwei unverbundene Stichproben (Paralleldesign)

Der Wilcoxon-Test für zwei unverbundene Stichproben wird auch als Mann-Whitney oder U-Test bezeichnet [33]. Auch für diesen Test ist die Nullhypothese als Gleichheit und die Alternativhypothese als Unterschied definiert [58]. Im Gegensatz zum Wilcoxon-Test für eine Stichprobe setzt diese Form weder Normalverteilung noch Symmetrie voraus [58]. Soll der Wilcoxon-Test für unverbundene Stichproben zur Anwendung gebracht werden, dann wird die Abweichung zwischen den beobachteten und den erwarteten Rangsummen betrachtet. Dazu werden zunächst Ränge vergeben, Rangsummen gebildet und die Testgröße bestimmt [33]. In diesem Fall gilt, dass das Ergebnis signifikant ist und die Nullhypothese abgelehnt wird, wenn der p-Wert kleiner oder gleich dem gewählten Signifikanzniveau ist [33, 58].

Wilcoxon-Test für zwei verbundene Stichproben (Wertepaare)

Der Wilcoxon-Test für zwei verbundene Stichproben stellt das Gegenstück zum entsprechenden t-Test dar und dient dem Vergleich der beiden Mediane. Auch in diesem Fall bezeichnet die Nullhypothese die Gleichheit und die Alternativhypothese den Unterschied [5, 58]. Wird der Wilcoxon-Test auf eine verbundene Versuchsanordnung angewendet, dann werden zunächst die Differenzen der Wertepaare gebildet und diese in Ränge eingeteilt. Anschließend werden für die positiven und die negativen Differenzen Rangsummen gebildet und die Prüfgröße ermittelt. Auch hier gilt, dass das Ergebnis signifikant wird für den

Fall, dass der p-Wert kleiner oder gleich dem gewählten Signifikanzniveau ist [33, 58]. Wichtig ist zu bemerken, dass die Fallzahl der Anzahl der Differenzen entspricht, die ungleich Null sind [33].

### *2.3.4 Chi-Quadrat-Test*

Mit Hilfe von Chi-Quadrat-Tests können Häufigkeitsunterschiede untersucht werden [59]. Chi-Quadrat-Tests dienen der Auswertung qualitativer, diskreter und kategorieller Merkmale bei mehr als zwei Parallelgruppen. Sie werden zum Vergleich der beobachteten Häufigkeiten mit den unter der Nullhypothese erwarteten Häufigkeiten eingesetzt. Es handelt sich um unabhängige Beobachtungen. Primär werden zwei Formen des Chi-Quadrat-Tests unterschieden.

Chi-Quadrat-Unabhängigkeitstest

Der Chi-Quadrat-Unabhängigkeitstest dient dem Vergleich zweier qualitativer Merkmale hinsichtlich ihrer stochastischen Unabhängigkeit [34, 59]. Die Nullhypothese definiert die Unabhängigkeit, die Alternativhypothese im Gegensatz dazu die Abhängigkeit [6, 59].

Chi-Quadrat- Homogenitätstest

Überprüft man mittels der Vier-Felder-Tafel, ob die Ausprägung der Zielgröße in beiden Gruppen gleich ist, so werden relative Häufigkeiten verglichen. Der hierfür angewendete Test, nach demselben Testverfahren wie der Chi-Quadrat-Unabhängigkeitstest, heißt Chi-Quadrat-Homogenitätstest [59].

Für die Ausprägungen werden absolute Häufigkeiten beobachtet und die unter Gültigkeit der Nullhypothese erwarteten absoluten Häufigkeiten berechnet [34]. Zur Bestimmung der Testgröße werden dazu die Differenzen beobachteter und erwartete Häufigkeiten quadriert und durch die jeweilige erwartete Häufigkeit dividiert. Die so entstehenden Chi-Quadrat-Anteile werden aufsummiert und bilden so die Testgröße, das Gesamt-Chi-Quadrat. Wird der Gesamt-Chi-Quadrat-Wert größer oder entspricht dem zugehörigen Tabellenwert, so wird die Entscheidung gegen die Nullhypothese getroffen und das Ergebnis als signifikant betrachtet [34].

Die Voraussetzungen für die Anwendung eines Chi-Quadrat-Tests sind wie folgt definiert: Es müssen absolute, nicht-kumulierte Häufigkeiten für die Berechnungen

verwendet werden [34]. Zudem müssen die beobachteten und die erwarteten Häufigkeiten groß genug sein. Dies bedeutet, dass die Zahl der beobachteten absoluten Häufigkeiten nicht Null sein darf (oder: $n \geq 2$) und die Zahl der erwarteten, absoluten Häufigkeiten $n \geq 5$ (oder: $n \geq 10$) sein muss [34, 59]. Ist dies nicht der Fall, sollte der Fisher-Exact-Test verwendet werden [59].

Im Rahmen der statistischen Fallzahlplanung kann die benötigte Fallzahl n bestimmt werden, wenn das Signifikanzniveau und der relevante Unterschied zwischen den Merkmalen, bei dem der Test mit einer vorgegebenen Power signifikant werden soll, gegeben sind [32].

### *2.3.5 LogRank-Test*

Um Überlebenszeiten unverbundener Stichproben vergleichen zu können, wird der LogRank-Test angewendet [59]. Er wird eingesetzt, wenn zwei oder mehr Gruppen, die sich hinsichtlich einer Einflussgröße unterscheiden, in Bezug auf die Überlebenskurven miteinander verglichen werden sollen [35, 59]. Dabei definiert die Nullhypothese den Fall, dass die Überlebenskurven aller Gruppen gleich sind [35]. Handelt es sich bei den erhobenen Daten um zensierte Daten, können auch diese mittels des LogRank-Tests ausgewertet werden [59].

Die Auswertung eines LogRank-Tests funktioniert folgendermaßen: Zuerst muss jedes Zeitintervall für sich betrachtet werden. Alle in einem Zeitintervall aufgetretenen Ereignisse werden auf sämtliche Gruppen verteilt und zwar abhängig von der Anzahl n innerhalb der Gruppen zu Beginn dieses Zeitintervalls. Es werden generell nur solche Intervalle betrachtet, in denen ein Ereignis stattfindet. Die Testgröße bildet sich schließlich als Summe der Ereignisse innerhalb aller Zeitintervalle. Ebenso wie für den Chi-Quadrat-Test gilt auch für den LogRank-Test, dass das Ergebnis signifikant wird, wenn die Testgröße größer oder gleich dem angegebenen Tabellenwert ist [34, 35].

Wird ein LogRank-Test bei der Auswertung der Ergebnisse angewendet, wird auf die Hazard-Ratio zurückgegriffen. Die Hazard-Funktion H(t) ist definiert als die Funktion des Sterberisikos. Die Hazard-Ratio ist definiert als das „Verhältnis zweier Hazardfunktionen“ ([29], S. 205) und gilt als die momentan betrachtete Sterberate bzw. als Maß für das relative Risiko zum Zeitpunkt (t). Durch Bestimmung der Hazard-Ratio für jeden beliebigen Zeitpunkt kann eine Prognose für eben diesen Zeitpunkt getroffen werden [36].

## 2.4 Biometrische Fallzahlplanung

Um eine biometrische Fallzahlplanung durchführen zu können, bedarf es einiger Vorinformationen. Solche Vorinformationen können aus der aktuellen Literatur oder im Rahmen sogenannter Pilotstudien generiert werden. So können Pilotstudien zum einen dazu dienen, den logistischen Aufwand eines Projektes abzuschätzen. Zum anderen können sie eine grobe Richtung vorgeben, in die die zu erwartenden Mittelwerte und Standardabweichungen des eigentlichen Tierversuches gehen werden. In manchen Fällen können sie auch einen Hinweis auf das Ergebnis des eigentlichen Projektes liefern [17, 25, 26, 42]. Pilotstudien sollten deshalb dann durchgeführt werden, wenn es keine Vorinformationen für das geplante Projekt gibt und somit eine biometrische Fallzahlplanung nicht möglich ist [42]. Allerdings gilt zu beachten, dass für Pilotstudien selbst keine statistische Fallzahlplanung möglich ist, da für sie noch keine Vorinformationen vorliegen oder das Ergebnis des Experiments nicht abschätzbar ist [17, 43]. Somit gilt, dass die Fallzahl für Pilotstudien auf der Erfahrung und Einschätzung des Testers beruht [17]. Allgemein vorgeschlagen werden für stetige Variablen fünf Tiere pro Gruppe, acht Tiere pro Gruppe für qualitative Zielgrößen und zehn Tiere pro Gruppe für den Fall einer Überlebenszeitanalyse [43]. Die Tierschutzkommission Tübingen, in deren Zuständigkeit die Universität Ulm fällt, nennt eine Tierzahl von „nicht mehr als 60, höchstens 80 Tiere“ [[43], S.369] für die Verwendung im Rahmen von Pilotstudien. Die Ergebnisse der Pilotstudie liefern dann die Vorinformationen für die Fallzahlplanung im Rahmen der eigentlichen Studie [17]. Generell gilt, dass die Fallzahl, die benötigt wird um ein signifikantes Ergebnis zu erhalten umso größer wird, je geringer der Unterschied ist, der untersucht werden soll und je größer die Variabilität in der Grundgesamtheit ist [17].

Um eine statistische Fallzahlplanung für ein Projekt durchführen zu können, müssen vorab folgende Parameter bekannt sein: Die Effektstärke, die Standardabweichung, das Signifikanzniveau, die Power und das Hypothesenpaar bzw. die Alternativhypothese [17, 25, 43]. Die Effektstärke muss quantitativ durch den Projektleiter angeben sein, die Standardabweichung wird aus der Pilotstudie oder der Literatur gewonnen, wohingegen Signifikanzniveau und Power durch allgemein anerkannte Konventionen bestimmt sind [17].

Das Ziel der biometrischen Fallzahlplanung ist es also, die Zahl der für ein Projekt benötigten Tiere zu berechnen und somit die Anzahl der Tiere zu optimieren [43], d.h. „das Optimum einer Fallzahl [ist] sowohl im Kontext minimal ausreichender als auch mindestens erforderlicher Fallzahl zu verstehen“ [[43], S.368]. Dies bedeutet, dass so viele Tiere wie nötig verwendet werden müssen, um den gewünschten Effekt mit statistischer Signifikanz nachweisen zu können, gleichzeitig sollte mit der berechneten Tierzahl der erwartete Effekt mit genügend Sicherheit gefunden werden [42, 43]. So ist es anhand des statistischen Verfahrens, mit dem die Ergebnisse ausgewertet werden sollen, möglich, die optimale Fallzahl zu berechnen [3, 26]. Obwohl für die Auswertung der Ergebnisse komplexe statistische Analysemethoden angewendet werden, ist es möglich die Fallzahlplanung mit einfacheren Modellen durchzuführen. Begründet ist dies darin, dass das Ziel der Fallzahlplanung lediglich darin besteht, die optimale Tierzahl zu berechnen, die meist gerundet wird, um eine angemessene Tierzahl für die entsprechende Versuchsplanung zu erhalten [17].

## 2.5 Statistikprogramm nQuery Advisor®

Die statistischen Fallzahlplanungen der dieser Dissertation zugrundeliegenden Tierversuchsgutachten des Instituts für Epidemiologie und Medizinische Biometrie der Universität Ulm der Jahre 2010 bis 2014 und die erneuten Berechnungen im Rahmen dieser Doktorarbeit wurden mit Hilfe des Programms *nQuery Advisor*® durchgeführt. Dieses Programm wird von der Firma Statistical Solutions mit Sitz in Cork, Irland und Boston/Massachusetts, USA vertrieben. Es bietet dem Benutzer die Möglichkeit, die für einen Versuch benötigte Fallzahl bei entsprechenden Vorgaben für das Signifikanzniveau, die Power, das Hypothesenpaar und testspezifische Bedingungen zu berechnen. Diese Berechnung der für einen Versuch benötigten Fallzahlen findet unter Anwendung des für die Auswertung vorgesehenen und somit im Voraus ausgewählten statistischen Testverfahrens statt [53]. Im Folgenden soll das Programm und seine Funktionsweise exemplarisch an den vier am häufigsten vorgefundenen Testverfahren – t-Test, Wilcoxon-Test, Chi-Quadrat-Test und LogRank-Test – mit Hilfe von entsprechenden Screenshots aus dem Programm *nQuery Advisor*® erklärt werden.

Startmenü

Nach Start des Programms und dem Öffnen einer neuen Datei, bietet sich dem Benutzer die Möglichkeit im Rahmen des Menüs *Study Goal and Design* aus einer Vielzahl statistischer Tests auszuwählen.

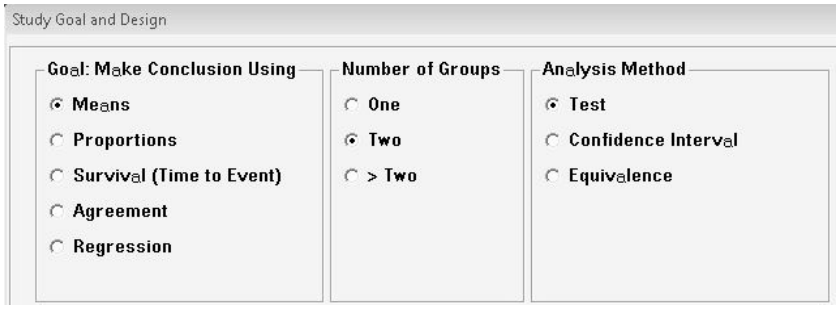


*Abbildung 4: Startmenü nQuery Advisor®*

Unter dem Menü-Punkt *Goal: Make Conclusion Using* wählt der Benutzer die der Berechnung zugrunde liegenden Parameter aus. Hier finden sich fünf Möglichkeiten: Mittelwerte (*Means*), Anteile (*Proportions*), Überlebenszeiten (*Survival/Time to Event*), Übereinstimmungen (*Agreement*) und Regressionen (*Regression*).

Mittels des Menü-Punktes *Number of Groups* wählt der Benutzer die Anzahl der Gruppen aus. Hier gibt es drei Optionen: eine Gruppe (*One*), zwei Gruppen (*Two*) oder mehr als zwei Gruppen (*>Two*).

In einem letzten Schritt muss noch die Analyse-Methode (*Analysis Method*) gewählt werden. Hierfür finden sich ebenfalls drei Möglichkeiten: Test auf Unterschiede (*Test*), Bestimmung des Konfidenzintervalls (*Confidence Interval*) und Äquivalenztest (*Equivalence*).

Sobald die gewünschten Einstellungen gewählt sind, erscheinen in einem neuen, darunterliegenden Feld sämtliche statistische Tests, die diese Eigenschaften erfüllen und von denen der Benutzer nun einen auswählen muss.

t-Test

Soll der späteren Datenanalyse ein t-Test zugrunde liegen, wird dieser auch für die Fallzahlplanung verwendet.

| Two group Satterthwaite t-test of equal means (unequal variances) (equal n's) | | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| Test significance level, α | | | | | |
| 1 or 2 sided test? | | | | | |
| Group 1 mean, $\mu_1$ | | | | | |
| Group 2 mean, $\mu_2$ | | | | | |
| Difference in means, $\mu_1 - \mu_2$ | | | | | |
| Group 1 standard deviation, $\sigma_1$ | | | | | |
| Group 2 standard deviation, $\sigma_2$ | | | | | |
| Power ( % ) | | | | | |
| n per group | | | | | |

*Abbildung 5: Maske für die Durchführung eines t-Tests mittels nQuery Advisor®*

Dazu wird die entsprechenden Maske des Programms *nQuery Advisor®* wie folgt befüllt: Signifikanzniveau α (*Test significance level, α*), ein- oder zweiseitiges Hypothesenpaar (*1 or 2 sided test?*), Mittelwert je Gruppe ($\mu_1$ *und* $\mu_2$), Differenz der Mittelwerte ($\mu_1 - \mu_2$), Standardabweichung je Gruppe ($\sigma_1$ *und* $\sigma_2$) und Power (*Power %*).

Sind alle diese Felder ausgefüllt, berechnet das Programm die benötigte Fallzahl n (*n per group*), um die vorgegebenen Zielwerte im Rahmen des Experiments zu erhalten.

Wilcoxon-Test

Soll bei der späteren Auswertung ein Rangsummentest im Sinne eines Wilcoxon-Tests zur Anwendung kommen, wird analog der Verwendung eines t-Tests verfahren.

| Wilcoxon (Mann-Whitney) rank-sum test that P(X<Y) = .5 (continuous outcome) | | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| Test significance level, α | | | | | | |
| 1 or 2 sided test? | | | | | | |
| $p_1 = P(X<Y)$ | | | | | | |
| Power ( % ) | | | | | | |
| n per group | | | | | | |

*Abbildung 6: Maske für die Durchführung eines Wilcoxon-Tests mittels nQuery Advisor®*

Nach Auswahl des Tests im Menü werden die folgenden Felder gefüllt: Signifikanzniveau α (*Test significance level, α*), ein- oder zweiseitiges Hypothesenpaar (*1 or 2 sided test?*), Wahrscheinlichkeit ($p_1 = P(X<Y)$) und Power (*Power %*).

Um die Wahrscheinlichkeit berechnen zu können, öffnet sich ein weiteres Fenster:

| Estimate of $p_1 = P(X<Y)$ from means and SD assuming normality | 1 | 2 | 3 |
|---|---|---|---|
| Group 1 mean, $\mu_1$ | | | |
| Group 2 mean, $\mu_2$ | | | |
| Difference in means, $\mu_1 - \mu_2$ | | | |
| Common standard deviation, $\sigma$ | | | |
| Effect size, $\delta = (\mu_1 - \mu_2) / \sigma$ | | | |
| $p_1 = P(X<Y)$ | | | |

*Abbildung 7: Maske zur Berechnung der Wahrscheinlichkeit im Rahmen der Durchführung eines Wilcoxon-Tests*

Hier werden folgende Felder ausgefüllt: Mittelwert je Gruppe ($\mu_1$ *und* $\mu_2$), Differenz der Mittelwerte ($\mu_1$ - $\mu_2$), gemeinsame Standardabweichung (*common standard deviation,* $\sigma$) und Effektgröße (*Effect size,* $\delta = (\mu_1 - \mu_2) / \sigma$). Der Wilcoxon-Test, der als nicht-parametrischer Test normalerweise zwei Mediane miteinander vergleicht, verlangt an dieser Stelle die Eingabe von Mittelwerten. Dies erscheint zunächst merkwerkwürdig. Grund dafür ist die im Hintergrund des Programms *nQuery Advisor*® ablaufende Approximation bzw. Näherung, die auf der Normalverteilung beruht und letztendlich zur Bestimmung der Fallzahl benötigt wird.

Zur Berechnung der gemeinsamen Standardabweichung öffnet sich ein drittes Fenster:

| Estimating common standard deviation by pooling $s_1$ and $s_2$ | 1 | 2 | 3 |
|---|---|---|---|
| $n_1$ | | | |
| $n_2$ | | | |
| Standard deviation of Group 1, $s_1$ | | | |
| Standard deviation of Group 2, $s_2$ | | | |
| Common standard deviation, $\sigma$ | | | |

*Abbildung 8: Maske zur Berechnung der gemeinsamen Standardabweichung (Wilcoxon-Test)*

Hier werden die folgenden Variablen eingetragen: Gruppengröße der beiden Gruppen ($n_1$ *und* $n_2$) und Standardabweichungen der Gruppen ($s_1$ *und* $s_2$).

Auf diese Weise kann die gemeinsame Standardabweichung der beiden Gruppen berechnet und zur Berechnung der Wahrscheinlichkeit herangezogen werden. Ist die Wahrscheinlichkeit bestimmt, kann die benötigte Fallzahl kalkuliert werden.

Chi-Quadrat-Test

Im Falle der Anwendung eines Chi-Quadrat-Tests zur Auswertung der Daten, wird dieser zur Bestimmung der Fallzahl herangezogen.

| Two group $\chi^2$ test of equal proportions (odds ratio = 1) (equal n's) | | |
|---|---|---|
| | 1 | 2 |
| Test significance level, $\alpha$ | | |
| 1 or 2 sided test? | | |
| Group 1 proportion, $\pi_1$ | | |
| Group 2 proportion, $\pi_2$ | | |
| Odds ratio, $\psi = \pi_2 (1 - \pi_1) / [\pi_1 (1 - \pi_2)]$ | | |
| Power ( % ) | | |
| n per group | | |

*Abbildung 9: Maske zur Durchführung eines Chi-Quadrat-Tests mittels nQuery Advisor®*

Dazu werden folgende Daten benötigt: Signifikanzniveau α (*Test significance level, α*), ein- oder zweiseitiges Hypothesenpaar (*1 or 2 sided test?*), Proportionen der beiden Gruppen (*$\pi_1$ und $\pi_2$*), Odds ratio (*$\psi = \pi_2 (1 - \pi_1) / [\pi_1 (1 - \pi_2)]$*) und Power (*Power %*).

Sind diese Felder ausgefüllt, kann die benötigte Fallzahl pro Gruppe berechnet werden.

LogRank-Test

Handelt es sich bei der Auswertung des Experiments um eine Überlebenszeitanalyse, so wird der LogRank-Test zur formalen Fallzahlplanung herangezogen.

| Log-rank test of survival in two groups followed for fixed time, constant hazard ratio | | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| Test significance level, $\alpha$ | | | | | |
| 1 or 2 sided test? | | | | | |
| Group 1 proportion $\pi_1$ at time t | | | | | |
| Group 2 proportion $\pi_2$ at time t | | | | | |
| Hazard ratio, $h=\ln(\pi_1) / \ln(\pi_2)$ | | | | | |
| Power ( % ) | | | | | |
| n per group | | | | | |
| Total number of events required, E | | | | | |

*Abbildung 10: Maske zur Durchführung eines LogRank-Tests mittels nQuery Advisor®*

Hier werden vom Benutzer folgende Felder befüllt: Signifikanzniveau α (*Test significance level, α*), ein- oder zweiseitiges Hypothesenpaar (*1 or 2 sided test?*), Proportionen der Gruppen zum Zeitpunkt t (*$\pi_1$ und $\pi_2$ at time t*), Hazard Ratio (*$h = \ln(\pi_1)/\ln(\pi_2)$*) und Power (*Power %*).

Sind diese Felder ausgefüllt, berechnet das Programm *nQuery Advisor*® die benötigte Fallzahl.

Es bleibt anzumerken, dass auch für die hier vorgestellten Tests jeweils nur die am häufigsten zum Einsatz gekommene Form besprochen wurde. Da das Programm *nQuery Advisor*® recht intuitiv funktioniert, soll auf die Erklärung sämtlicher anderer Möglichkeiten verzichtet werden.

## 2.6 Berechnungen

Zur Beantwortung der Fragestellung dieser Dissertation sind verschiedene Szenarien durchgespielt worden, die im Folgenden vorgestellt und erklärt werden sollen.

Für die Berechnung der originalen Tierversuchsgutachten fand eine primäre Fallzahlberechnung mit Hilfe des Programms *nQuery Advisor*® statt. Dazu wurden aus den Angaben der Antragssteller die benötigten Daten entnommen und in die entsprechende Maske eingegeben. Anschließend wurde mithilfe dieser ersten Fallzahl $n_1$, je nach Fragestellung, die Fallzahl $n_2$ bestimmt, die für die Durchführung des jeweiligen Versuches von Nöten ist.

Bei der Berechnung von $n_2$ wurden folgende Besonderheiten der einzelnen Experimente berücksichtigt:

Handelte es sich um Versuche mit mehreren Gruppen, so mussten die Fallzahlen für jede Gruppe getrennt berechnet und am Ende aufsummiert werden, um die Gesamtfallzahl zu erhalten.

Lag ein Experiment mit Spender- und Empfängertieren vor, mussten auch hier sämtliche Tiere in der Gesamtfallzahl berücksichtigt werden. Hier verhielt es sich häufig so, dass lediglich eine Berechnung der Empfängertiere stattfand und die Anzahl der Spendertiere in solchen Fällen in einem bestimmten, vorgegebenen Verhältnis zu den Empfängertieren berechnet wurde. Benötigte man beispielsweise für drei Empfängertiere ein Spendertier, lies sich die Anzahl dieser über das vorgegebene Verhältnis von 3:1 berechnen. In solchen Fällen musste, damit eine mathematische Rechnung sinnvoll möglich war, die Anzahl der Empfängertiere auf eine durch z.B. drei teilbare Anzahl aufgerundet werden.

In manchen Fällen war vom Antragssteller vorgegeben worden, dass die Fallzahl eine, beispielsweise durch vier teilbare Zahl ist, da dies für den Versuchsaufbau

so benötigt wurde. Aus diesem Grund wurden die berechneten Fallzahlen entsprechend aufgerundet.

Wurden bei der Berechnung der primären Fallzahl $n_1$ durch Eingabe unterschiedlicher Mittelwerte große Unterschiede erzielt, fand eine auf Erfahrungswerten basierende Annahme der mittleren Fallzahl statt.

In fast allen Fällen wurden Ersatztiere mit einberechnet, falls Tiere im Laufe einer Versuchsreihe vorzeitig versterben. Die Anzahl der Ersatztiere betrug in der Regel einen gewissen Prozentsatz (zwischen 5% und 20%) der benötigten Fallzahl. Die Anzahl der benötigten Ersatztiere wurde dabei durch den Antragssteller festgesetzt und bei der Berechnung der für den Versuch nötigen Gesamttierfallzahl berücksichtigt.

Für den Fall, dass die berechnete Fallzahl nicht zur Menge der natürlichen Zahlen gehörte, wurde die Fallzahl nach den gültigen mathematischen Regel auf- bzw. abgerundet.

Um die Vergleichbarkeit der im Rahmen der durchgeführten Re-Kalkulationen neu berechneten Fallzahlen zu gewährleisten, wurde in den oben geschilderten Fällen exakt so verfahren wie in den Originalgutachten. Dazu wurden die für die Neuberechnungen benötigten Daten den originalen Tierversuchsgutachten entnommen. Anschließend wurde zu jedem der 112 berücksichtigten Tierversuchsgutachten insgesamt 18 Berechnungen mittels des Programms *nQuery Advisor*® durchgeführt. Dazu wurden alle 17 neuen Bedingungen sowie die Ausgangsbedingung simuliert und die durch das Programm nQuery Advisor® berechnete Fallzahl dazu verwendet, um die endgültige Tierversuchsfallzahl zu berechnen. Die folgende Tabelle gibt einen Überblick über die genannten 18 Szenarien, die im Rahmen der Re-Kalkulationen berücksichtigt wurden.

*Tabelle 4: Darstellung der jeweils durchgeführten 18 Berechnungen und den entsprechenden Bedingungen*

| | | Power 1-β in % | | | | | |
|---|---|---|---|---|---|---|---|
| | | 70 | 80 | 90 | 70 | 80 | 90 |
| Signifikanz-niveau α in % | 1 | | | | | | |
| | 5 | | | | | | |
| | 10 | | | | | | |
| | | 1-seitig | | | 2-seitig | | |
| | | Hypothesenpaar | | | | | |

In einer ersten Tabelle wurden die neu berechneten Fallzahlen $n_1$ eingetragen. Da in der Regel einer der neu berechneten Werte mit dem des Originalgutachtens übereinstimmte, fand bereits hier eine erste Überprüfung bezüglich möglicher Fehler statt. Zudem muss die Fallzahl für die Kombination α = 5%, 1-β = 80% bei einseitiger Testung mit der Fallzahl für die Kombination α = 10%, 1-β = 80% bei zweiseitigem Hypothesenpaar übereinstimmen. Dies bietet die Möglichkeit einer zweiten Kontrolle, ob sich nicht bereits zu Beginn der Berechnung Fehler eingeschlichen haben. Im nächsten Schritt wurde anhand des vorliegenden Gutachtens mit den berechneten Fallzahlen $n_1$ genau wie dort verfahren, um die endgültigen Fallzahlen für die einzelnen Gruppen oder Versuchsteile zu erhalten. Hier fanden beispielsweise Vorgaben des Antragsstellers bezüglich Spender- und Empfängertieren oder der Teilbarkeit der Tierversuchszahl Berücksichtigung. Anschließend sind mittels des vorgegebenen Schlüssels die Ersatztiere anhand der bis dahin vorliegenden Gesamtzahl berechnet worden. Zum Schluss wurden die Fallzahlen für die einzelnen Gruppen und/oder Versuchsteile und die Zahl der Ersatztiere summiert, um die Gesamtfallzahl zu erhalten.

Wie oben beschrieben wurde mit allen 18 berechneten Werten verfahren, damit aufgezeigt werden konnte, wie sich eine Veränderung der statistischen Parameter Signifikanzniveau, Power und Hypothesenpaar auf die Gesamtfallzahl auswirkt.

### *2.6.1 Ausgangskonstellationen*

Die im Folgenden genannten vier Ausgangskonstellationen wurden im Rahmen der Neuberechnungen berücksichtigt und anschließend ausgewertet:

1. α = 5% 1-β = 80% Hypothesenpaar = 2-seitig
2. α = 2,5% 1-β = 80% Hypothesenpaar = 1-seitig
3. α = 5% 1-β = 80% Hypothesenpaar = 1-seitig
4. α = 1,67% 1-β = 80% Hypothesenpaar = 2-seitig

Diese vier Ausgangskonstellationen sind diejenigen, die den berücksichtigten Tierversuchsgutachten zu Grunde lagen, und deshalb zur Endauswertung herangezogen wurden.

Anzumerken ist, dass Konstellation 1 mit Konstellation 2 „gleichzusetzen“ ist und sowohl für die Neuberechnung als auch für die Auswertung dieselbe Tabelle

verwendet wurde wie für Ausgangskonstellation 1 ([41, 56], siehe auch: 2.2.4 Hypothesenpaare - Einfluss des Hypothesenpaares auf das Signifikanzniveau).

Betrachtet man die Verteilung der 112 in die Auswertung eingegangenen Tierversuchsgutachten auf die vier zugrunde liegenden Ausgangskonstellationen zeigt sich folgendes Bild:

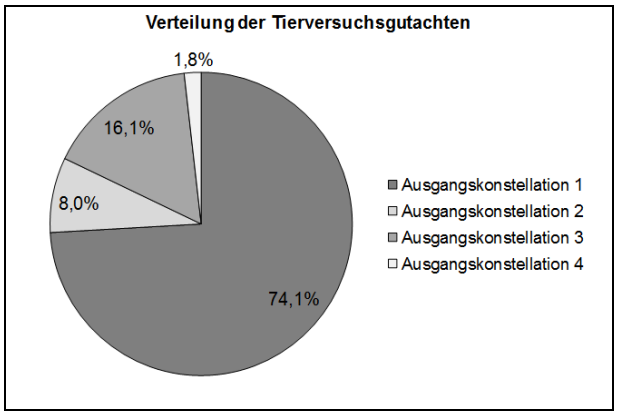


*Abbildung 11: Prozentuale Verteilung der betrachteten Tierversuchsgutachten auf die vier Ausgangskonstellationen*

Aus der Abbildung wird ersichtlich, dass nahezu dreiviertel aller berücksichtigten Tierversuchsgutachten der Ausgangskonstellation 1 zuzuordnen sind. Das entspricht, in Zahlen ausgedrückt, 83 Gutachten. Lediglich 9 Gutachten, entsprechend 8,0%, lag die Ausgangskonstellation 2 zu Grunde. Mit 18 Tierversuchsgutachten beruhen 16,1% der Gutachten auf Ausgangskonstellation 3 und nur bei 2 Gutachten, und damit 1,8%, lag die Ausgangskonstellation 4 vor.

### *2.6.2 Auswertung*

In der abschließenden Auswertung werden jeweils die neu berechneten Fallzahlen mit der entsprechenden Ausgangskonstellation verglichen und die Veränderung der Fallzahl in Prozent berechnet. Dies wurde folgendermaßen durchgeführt:

<u>Vergleich der Jahre – sortiert nach Ausgangskonstellation und betrachtet für die Summe der Jahre</u>

Bei dieser Berechnung werden sämtliche Gutachten der fünf berücksichtigten Jahre, denen dieselbe Ausgangskonstellation zugrunde liegt, gemeinsam betrachtet. Dabei wird die Änderung der Fallzahl für jede der 17 neu betrachteten Konstellationen in Bezug zur jeweiligen, den entsprechenden Gutachten zugrunde liegenden Ausgangskonstellation gesetzt und verglichen. Diese Auswertung erfolgt getrennt für die vier oben genannten Konstellationen.

Vergleich der Fallzahlen – sortiert nach Ausgangskonstellation und Jahr

Alle Gutachten eines Jahres, die dieselbe Konstellation aufweisen, werden zusammengefasst und gemeinsam verglichen. Dazu werden alle 17 neu betrachteten Szenarien bzw. die daraus resultierenden Fallzahlen mit der Gesamtfallzahl der entsprechenden Originalgutachten und damit der entsprechenden Ausgangskonstellation verglichen. Dies erfolgt getrennt für jede der vier oben erwähnten Konstellationen.

## 2.7 Nicht berücksichtigte Gutachten

Wie bereits oben erwähnt, konnten insgesamt 52 Gutachten nicht in der Endberechnung berücksichtigt werden.

Bei 27 dieser Gutachten handelte es sich um Pilotstudien, die aufgrund dessen, dass bereits für das Originalgutachten keine statistische Fallzahlplanung stattgefunden hat, nicht neu berechnet werden konnten.

Die anderen 25 Gutachten wiesen andere Gründe auf, warum sie nicht in die Endauswertung mit einbezogen werden konnten. Diese Gründe werden im Folgenden kurz erläutert.

Bei 13 Gutachten waren entweder die komplette Berechnung oder Teile der Berechnung mit dem Programm *nQuery Advisor*® nicht vorhanden, sodass die ursprüngliche Fallzahlplanung nicht nachvollzogen werden konnte und dementsprechend auch keine neue Berechnung möglich war.

Bei einem Gutachten wurde die Variable der statistischen Power, die im Rahmen der Neuberechnung variiert wurde, nicht in der Ausgangsberechnung berücksichtigt, sodass dieses Gutachten nicht weiter bearbeitet werden konnte.

Bei fünf der Gutachten fand bereits im Rahmen der Ausgangsberechnung keine Testung hinsichtlich ein- und zweiseitigem Hypothesenpaar statt, sodass diese Gutachten ebenfalls keinen Eingang in die Endberechnung fanden. Die Begründung hierfür findet sich in der Tatsache, dass die Ausrichtung des Hypothesenpaares eine der Variablen für die Neuberechnung und auch die Endauswertung darstellt. Da die Ausrichtung des Hypothesenpaares bei diesen Gutachten jedoch nicht variabel war, konnten diese Gutachten, nicht weiter verfolgt werden.

Ein weiteres Gutachten konnte auf Grund seiner Eigenschaft als reine Stellungnahme ebenfalls nicht berücksichtigt werden, da hier keine statistische Fallzahlplanung von Seiten des Instituts für Epidemiologie und Medizinische Biometrie stattgefunden hat.

Aufgrund großer Komplexität und der dadurch bedingten Gefahr von Fehlern in der Berechnung, wurden drei weitere Gutachten nicht bearbeitet.

Fünf weitere Gutachten sind aufgrund dessen, dass der im jeweiligen Antrag geschilderte Versuch in Bezug zur ökologischen oder tiermedizinischen Forschung stand, nicht berücksichtigt worden. Bei zwei dieser Gutachten handelte es sich gleichzeitig um Pilotstudien, sodass diese bereits dort berücksichtigt wurden. Bei einem dieser Gutachten fand zudem keine Berücksichtigung der Power statt, sodass dieses bereits dort mitgezählt wurde. In der Summe finden sich somit nur zwei dieser Gutachten in der Zählung der 25 Gutachten, die aus einem anderen Grund als ihrer Eigenschaft als Pilotstudie von der Berücksichtigung ausscheiden.

Als weitere mögliche Fehlerquelle wären Flüchtigkeitsfehler zu nennen. So ist es durchaus möglich, dass sich Rechen-, Übertragungs- oder Tippfehler eingeschlichen haben, obwohl sämtliche Daten mehrmals kontrolliert wurden. Zudem ist es denkbar, dass bereits die ursprünglichen Originalgutachten fehlerhaft sind. Auch hier sind Rechen-, Übertragungs- oder Tippfehler nicht sicher auszuschließen. Dies konnte allerdings im Rahmen dieser Arbeit nicht kontrolliert werden.

# 3 Ergebnisse

Im folgenden Teil sollen die Ergebnisse der 17 durchgeführten Berechnungen und die Ausgangsfallzahl in Form von Tabellen und kurzen Zusammenfassungen dargestellt werden.

## 3.1 Vergleich der Fallzahlen – sortiert nach Ausgangskonstellation und betrachtet für die Summe der Jahre 2010 - 2014

Im folgenden Teil werden alle vier Ausgangskonstellationen summiert über alle fünf betrachteten Jahre vorgestellt.

### *3.1.1 Ausgangskonstellation 1*

Der Berechnung liegt folgende Ausgangskonstellation zu Grunde:

α = 5%  1-β = 80%  Hypothesenpaar = 2-seitig

Im Rahmen der Re-Kalkulationen ergeben sich folgende Fallzahlen, die einer Änderung der Fallzahl (in %) im Vergleich zur Ausgangskonstellation entsprechen:

*Tabelle 5: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 5%, 1-β = 80%, Hypothesenpaar = 2-seitig für die Jahre 2010-2014 (Tierversuchsgenehmigungen Universität Ulm)*

| | | Hypothesenpaar | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1-seitig | | | 2-seitig | | |
| Signifikanz-niveau α in % | 1 | 47.492 (+5,1) | 52.665 (+16,5) | 60.502 (+33,9) | 52.697 (+16,6) | 57.982 (+28,3) | 66.761 (+47,7) |
| | 5 | 34.887 (-22,8) | 38.930 (-13,9) | 45.440 (+0,6) | 40.092 (-11,3) | **45.192** | 52.394 (+15,9) |
| | 10 | 30.953 (-31,5) | 33.746 (-25,3) | 38.835 (-14,1) | 34.556 (-23,5) | 38.930 (-13,9) | 45.468 (+0,6) |
| | | 70 | 80 | 90 | 70 | 80 | 90 |
| | | Power 1-β in % | | | | | |

Bei Ausgangskonstellation 1 handelt es sich um die am häufigsten verwendete und aufgrund der oben genannten Voraussetzungen für die Durchführung von Tierversuchen im Rahmen der medizinischen Grundlagenforschung auch die wichtigste der vier betrachteten Ausgangskonstellationen. Aus diesem Grund wird bereits an dieser Stelle genauer auf die Ergebnisse eingegangen. Bei der Betrachtung der Ergebnisse der Berechnungen im Rahmen dieser

Ausgangskonstellation über alle fünf berücksichtigten Jahre hinweg, zeigt sich bei getrennter Untersuchung der drei verschiedenen „Stellparameter“ Signifikanzniveau, Power und Hypothesenpaar folgendes:

Wird eine Veränderung der Power durchgeführt, so kann dies zu einer Erhöhung oder Erniedrigung der Fallzahl führen. Genauer gesagt führt eine alleinige Erhöhung der Power auf 90% zu einem Anstieg der Tierfallzahl um 15,9%. Eine alleinige Verringerung der Power auf 70% bedingt eine Reduktion der Fallzahl von 11,3%.

Bei der Betrachtung der Hypothesenpaare ist ähnliches zu beobachten. So zieht eine Veränderung des Hypothesenpaares von zweiseitig auf einseitig eine Verminderung der Fallzahl von 13,9% nach sich.

Wird das Signifikanzniveau, bei ansonsten gleichbleibenden Annahmen für Power und Hypothesenpaar, verändert, so lässt sich bei einem strengeren Signifikanzniveau von 1% ein Anstieg der Fallzahl um 28,3% beobachten. Eine Änderung des Signifikanzniveaus auf das liberalere Niveau von 10% hingegen führt zu einer Verringerung der Fallzahl um 13,9%, was einer Summe von 6.262 Tieren entspricht.

Bei kombinierter Veränderung dieser drei „Stellparameter“ kann bei der Kombination aus $\alpha = 1\%$, $1-\beta = 90\%$ und zweiseitigem Hypothesenpaar, was der „strengsten“ möglichen Kombination entspricht, eine Erhöhung der Tierfallzahl von fast 50% (genauer: 47,7%) beobachtet werden. Im Gegenzug führt die „liberalste“ Kombination aus $\alpha = 10\%$, $1-\beta = 70\%$ und einseitigem Hypothesenpaar gegensinnig zu einer Reduktion der Fallzahl um 31,5%.

Zusammenfassend wird gezeigt, dass eine Veränderung der Ausgangsbedingungen durchaus große Veränderungen in der beobachteten Tierfallzahl zur Folge hat. Dieses Prinzip zieht sich durch alle im Folgenden vorgestellten Berechnungen, weshalb dort nur noch auf die Veränderung der Parameter hin zu liberaleren Annahmen eingegangen wird.

### *3.1.2 Ausgangskonstellation 2*

Der Berechnung liegt folgende Ausgangskonstellation zu Grunde:

α = 2,5%  1-β = 80%  Hypothesenpaar = 1-seitig

Im Rahmen der Re-Kalkulationen ergeben sich folgende Fallzahlen, die einer Änderung der Fallzahl (in %) im Vergleich zur Ausgangskonstellation entsprechen:

*Tabelle 6: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 2,5%, 1-β = 80%, Hypothesenpaar = 1-seitig für die Jahre 2010-2014 (Tierversuchsgenehmigungen Universität Ulm)*

| | | Hypothesenpaar | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1-seitig | | | 2-seitig | | |
| Signifikanz-niveau α in % | 1 | 7.333 (+11,3) | 7.912 (+20,1) | 9.430 (+43,1) | 8.462 (+28,4) | 9.305 (+41,2) | 10.643 (+61,5) |
| | 5 | 4.850 (-26,4) | 5.707 (-13,4) | 6.316 (-4,1) | 5.963 (-9,5) | **6.589** | 7.785 (+18,2) |
| | 10 | 4.061 (-38,4) | 4.645 (-29,5) | 5.212 (-20,9) | 4.850 (-26,4) | 5.707 (-13,4) | 6.449 (-2,1) |
| | | 70 | 80 | 90 | 70 | 80 | 90 |
| | | Power 1-β in % | | | | | |

Für die Auswertung der Ausgangskonstellation 2 kann dieselbe Tabelle verwendet werden wie für die Auswertung der Ausgangskonstellation 1. Dies hat Gültigkeit, da ein Signifikanzniveau α von 2,5% bei einseitigem Hypothesenpaar einem Signifikanzniveau von 5% bei zweiseitigem Hypothesenpaar entspricht. Die Erklärung hierfür findet sich unter 2.2.4 Hypothesenpaare – Einfluss des Hypothesenpaares auf das Signifikanzniveau und 2.6.1 Ausgangskonstellationen. Bei einer Gesamtbetrachtung der Ausgangskonstellation 2 ergibt sich bei einer Erhöhung des Signifikanzniveaus von 2,5% auf 5%, was gleichzusetzen ist mit einer Erhöhung des Signifikanzniveaus von 5% auf 10%, eine Reduktion der Gesamttierfallzahl um 13,4%. Dies entspricht einer Anzahl von 882 Tieren. Auch im Rahmen der Ausgangskonstellation 2 ist eine Betrachtung hinsichtlich der beiden weiteren Stellparameter Power und Hypothesenpaar möglich. In Einklang mit den, bei der Betrachtung der Ausgangskonstellation 1 gemachten Beobachtungen, kann auch hier bei Erniedrigung der Power auf 70% eine Reduktion der Tierfallzahl festgestellt werden. Es handelt sich hier um eine Verminderung der Fallzahl um 9,5%. Wird das Hypothesenpaar von einer

zweiseitigen hin zu einer einseitigen Testung verändert, so kann die Gesamtfallzahl um 13,4% erniedrigt werden.

### *3.1.3 Ausgangskonstellation 3*

Der Berechnung liegt folgende Ausgangskonstellation zu Grunde:

α = 5%  1-β = 80%  Hypothesenpaar = 1-seitig

Im Rahmen der Re-Kalkulationen ergeben sich folgende Fallzahlen, die einer Änderung der Fallzahl (in %) im Vergleich zur Ausgangskonstellation entsprechen:

*Tabelle 7: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 5%, 1-β = 80%, Hypothesenpaar = 1-seitig für die Jahre 2010-2014 (Tierversuchsgenehmigungen Universität Ulm)*

| | | Hypothesenpaar | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1-seitig | | | 2-seitig | | |
| Signifikanz-niveau α in % | 1 | 10.249 (+23,0) | 11.767 (+41,2) | 13.734 (+64,9) | 11.730 (+40,8) | 12.851 (+54,3) | 15.156 (+81,9) |
| | 5 | 7.642 (-8,3) | **8.331** | 10.032 (+20,4) | 8.591 (+3,1) | 9.839 (+18,1) | 11.552 (+38,7) |
| | 10 | 6.645 (-20,2) | 7.077 (-15,1) | 8.815 (+5,8) | 7.642 (-8,3) | 8.331 (±0) | 9.972 (+19,7) |
| | | 70 | 80 | 90 | 70 | 80 | 90 |
| | | Power 1-β in % | | | | | |

Für die oben genannte Ausgangskonstellation ergibt sich, bei Betrachtung aller fünf Jahre gemeinsam, folgendes: Bei einer Erhöhung des Signifikanzniveaus um 5% auf 10% kann eine Verminderung der Tierfallzahl um 1.254 Tiere bzw. um 15,1% beobachtet werden. Wie bereits bei den beiden vorhergehenden Ausgangskonstellationen gezeigt, führt auch bei Betrachtung der Ausgangskonstellation 3 eine Veränderung der Power auf 70% zu einer Reduktion der Fallzahl um 8,3%. Da es sich bei Ausgangskonstellation 3 bereits von Beginn an um eine Testung zu einem einseitigen Hypothesenpaar handelt, entfällt diese Betrachtung hier.

### *3.1.4 Ausgangskonstellation 4*

Der Berechnung liegt folgende Ausgangskonstellation zu Grunde:

α = 1,67%  1-β = 80%  Hypothesenpaar = 2-seitig

Im Rahmen der Re-Kalkulationen ergeben sich folgende Fallzahlen, die einer Änderung der Fallzahl (in %) im Vergleich zur Ausgangskonstellation entsprechen:

*Tabelle 8: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 1,67%, 1-β = 80%, Hypothesenpaar = 2-seitig für die Jahre 2010-2014 (Tierversuchsgenehmigungen Universität Ulm)*

| | | Hypothesenpaar | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1-seitig | | | 2-seitig | | |
| Signifikanz-niveau α in % | 1,67 | 112 (-21,1) | 121 (-14,8) | 142 (±0) | 121 (-14,8) | **142** | 166 (+16,9) |
| | 5 | 94 (-33,8) | 103 (-27,5) | 112 (-21,1) | 103 (-27,5) | 112 (-21,1) | 136 (-4,2) |
| | 10 | 73 (-48,6) | 82 (-42,3) | 103 (-27,5) | 76 (-46,5) | 103 (-27,5) | 112 (-21,1) |
| | | 70 | 80 | 90 | 70 | 80 | 90 |
| | | Power 1-β in % | | | | | |

Für eine Erhöhung des Signifikanzniveaus α von 1,67% auf 5%, bei ansonsten gleichbleibenden Bedingungen, ergibt sich eine Reduktion der Tierfallzahl um 21,1%, entsprechend einer Anzahl von 30 Tieren. Dabei gilt es allerdings zu bedenken, dass in diesem Fall keine Adjustierung des Signifikanzniveaus mehr stattfindet. Im Vergleich zur Ausgangskonstellation 3, ist bei Ausgangskonstellation 4 eine Betrachtung der Power und des Hypothesenpaares möglich. Analog zu den, bei den anderen Ausgangskonstellationen gemachten Beobachtungen, führt auch im Rahmen der Ausgangskonstellation 4 eine Liberalisierung der Power auf 70% zu einer Reduktion der Fallzahl um genau 14,8%, was 21 Tieren entspricht. Auch eine Veränderung des Hypothesenpaares von zweiseitig zu einseitig zieht eine Verminderung der Fallzahl nach sich. Es handelt sich hier um eine Ersparnis von 14,8% der Tiere, gleichzusetzen mit einer Zahl von 21 Tieren.

## 3.2 Vergleich der Fallzahlen – sortiert nach Ausgangskonstellation und Jahr

Zunächst sollen die vier verschiedenen Ausgangskonstellationen sortiert nach den Jahren 2010 bis 2014 vorgestellt werden. Für die am häufigsten vorkommende Ausgangskonstellation, Konstellation 1 mit α = 5%, 1-β = 80%, Hypothesenpaar = 2-seitig, werden die Ergebnisse in Tabellenform dargestellt. Die entsprechenden Tabellen für alle anderen Szenarien finden sich im Anhang dieser Arbeit.

### *3.2.1 Ausgangskonstellation 1*

Der Berechnung liegt folgende Ausgangskonstellation zu Grunde:

α = 5%  1-β = 80%  Hypothesenpaar = 2-seitig

Jahr 2010

Im Rahmen der Re-Kalkulationen ergeben sich folgende Fallzahlen, die einer Änderung der Fallzahl (in %) im Vergleich zur Ausgangskonstellation entsprechen:

*Tabelle 9: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 5%, 1-β = 80%, Hypothesenpaar = 2-seitig für das Jahr 2010 (Tierversuchsgenehmigungen Universität Ulm)*

| | | Power 1-β in % | | | | | |
|---|---|---|---|---|---|---|---|
| | | 70 | 80 | 90 | 70 | 80 | 90 |
| Signifikanz-niveau α in % | 1 | 5.878 (+5,6) | 6.756 (+21,4) | 7.932 (+42,6) | 6.607 (+18,8) | 7.548 (+35,7) | 8.915 (+60,2) |
| | 5 | 4.107 (-26,2) | 4.724 (-15,1) | 5.680 (+2,1) | 4.910 (-11,8) | **5.564** | 6.813 (+22,5) |
| | 10 | 3.382 (-39,2) | 3.907 (-29,8) | 4.698 (-15,6) | 4.107 (-26,2) | 4.724 (-15,1) | 5.681 (+2,1) |
| | | 1-seitig | | | 2-seitig | | |
| | | Hypothesenpaar | | | | | |

Die Auswertung der Gutachten des Jahres 2010, denen die oben genannte Konstellation zu Grunde liegt, ergibt eine mögliche Reduktion der Fallzahl um 15,1% für den Fall, dass das Signifikanzniveau α mit 10% angenommen wird.

Jahr 2011

Im Rahmen der Re-Kalkulationen ergeben sich folgende Fallzahlen, die einer Änderung der Fallzahl (in %) im Vergleich zur Ausgangskonstellation entsprechen:

*Tabelle 10: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 5%, 1-β = 80%, Hypothesenpaar = 2-seitig für das Jahr 2011 (Tierversuchsgenehmigungen Universität Ulm)*

| | | Power 1-β in % | | | | | |
|---|---|---|---|---|---|---|---|
| | | 70 | 80 | 90 | 70 | 80 | 90 |
| Signifikanz-niveau α in % | 1 | 8.471 (+7,8) | 9.080 (+15,6) | 10.480 (+33,4) | 9.117 (+16,0) | 10.183 (+29,6) | 11.394 (+45,0) |
| | 5 | 6.242 (-20,6) | 6.780 (-13,7) | 7.886 (+0,4) | 7.173 (-8,7) | **7.858** | 9.129 (+16,2) |
| | 10 | 5.517 (-29,8) | 5.965 (-24,1) | 6.747 (-14,1) | 5.911 (-24,8) | 6.780 (-13,7) | 7.886 (+0,4) |
| | | 1-seitig | | | 2-seitig | | |
| | | Hypothesenpaar | | | | | |

Für das Jahr 2011 ergibt sich bei einem Signifikanzniveau α von 10% eine Reduktion der Fallzahl um 1.078 Tiere, was 13,7% entspricht.

Jahr 2012

Im Rahmen der Re-Kalkulationen ergeben sich folgende Fallzahlen, die einer Änderung der Fallzahl (in %) im Vergleich zur Ausgangskonstellation entsprechen:

*Tabelle 11: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 5%, 1-β = 80%, Hypothesenpaar = 2-seitig für das Jahr 2012 (Tierversuchsgenehmigungen Universität Ulm)*

| | | Power 1-β in % | | | | | |
|---|---|---|---|---|---|---|---|
| | | 70 | 80 | 90 | 70 | 80 | 90 |
| Signifikanz-niveau α in % | 1 | 12.787 (+6,4) | 14.183 (+18,0) | 16.315 (+35,7) | 14.226 (+18,4) | 15.441 (+28,5) | 18.016 (+49,9) |
| | 5 | 9.581 (-20,3) | 10.614 (-11,7) | 12.702 (+5,7) | 10.723 (-10,8) | **12.019** | 14.401 (+19,8) |
| | 10 | 8.518 (-29,1) | 9.337 (-22,3) | 11.054 (-8,0) | 9.581 (-20,3) | 10.614 (-11,7) | 12.702 (+5,7) |
| | | 1-seitig | | | 2-seitig | | |
| | | Hypothesenpaar | | | | | |

Durch die Veränderung des Signifikanzniveaus α auf 10% kann die Fallzahl um 11,7% reduziert werden.

Jahr 2013

Im Rahmen der Re-Kalkulationen ergeben sich folgende Fallzahlen, die einer Änderung der Fallzahl (in %) im Vergleich zur Ausgangskonstellation entsprechen:

*Tabelle 12: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 5%, 1-β = 80%, Hypothesenpaar = 2-seitig für das Jahr 2013 (Tierversuchsgenehmigungen Universität Ulm)*

| | | Power 1-β in % | | | | | |
|---|---|---|---|---|---|---|---|
| | | 70 | 80 | 90 | 70 | 80 | 90 |
| Signifikanz-niveau α in % | 1 | 3.835 (+2,8) | 4.017 (+7,6) | 4.414 (+18,3) | 4.017 (+7,6) | 4.324 (+15,7) | 4.746 (+27,2) |
| | 5 | 3.266 (-12,5) | 3.496 (-6,3) | 3.739 (+0,2) | 3.472 (-7,0) | **3.732** | 4.013 (+7,5) |
| | 10 | 2.997 (-19,7) | 3.163 (-15,2) | 3.417 (-8,4) | 3.266 (-12,5) | 3.496 (-6,3) | 3.750 (+0,5) |
| | | 1-seitig | | | 2-seitig | | |
| | | Hypothesenpaar | | | | | |

Im Jahr 2013 ergibt sich bei einem Signifikanzniveau von α = 10% ein Einsparungspotenzial von 6,32%.

Jahr 2014

Im Rahmen der Re-Kalkulationen ergeben sich folgende Fallzahlen, die einer Änderung der Fallzahl (in %) im Vergleich zur Ausgangskonstellation entsprechen:

*Tabelle 13: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 5%, 1-β = 80%, Hypothesenpaar = 2-seitig für das Jahr 2014 (Tierversuchsgenehmigungen Universität Ulm)*

| | | Power 1-β in % | | | | | |
|---|---|---|---|---|---|---|---|
| | | 70 | 80 | 90 | 70 | 80 | 90 |
| Signifikanz-niveau α in % | 1 | 16.521 (+3,1) | 18.629 (+16,3) | 21.361 (+33,4) | 18.730 (+16,9) | 20.486 (+27,9) | 23.690 (+47,9) |
| | 5 | 11.691 (-27,0) | 13.316 (-16,9) | 15.433 (-3,7) | 13.814 (-13,8) | **16.019** | 18.038 (+12,6) |
| | 10 | 10.539 (-34,2) | 11.374 (-29,0) | 12.919 (-19,4) | 11.691 (-27,0) | 13.316 (-16,9) | 15.449 (-3,6) |
| | | 1-seitig | | | 2-seitig | | |
| | | Hypothesenpaar | | | | | |

Bei einem Signifikanzniveau von 10% ergibt sich für das Jahr eine Reduktion der Fallzahl um 16,9%.

### *3.2.2 Ausgangskonstellationen 2, 3 und 4*

Aufgrund dessen, dass die Ausgangskonstellationen 2, 3 und 4 insgesamt deutlich seltener waren als die Ausgangskonstellation 1 und die ihnen zugrunde liegenden Vorgaben insgesamt im Rahmen des Tierversuchswesens in der medizinischen Grundlagenforschung eine eher untergeordnete Rolle spielen, finden sich die zugehörigen Tabellen mit allen Zahlen im Detail im Anhang dieser Arbeit. Die im Rahmen dieser Arbeit erhaltenen Ergebnisse werden im Folgenden zusammenfassend dargestellt.

Ausgangskonstellation 2

Der Berechnung liegt folgende Ausgangskonstellation zu Grunde:

α = 2,5%  1-β = 80%  Hypothesenpaar = 1-seitig

Für das Jahr 2010 wurde keines der Tierversuchsgutachten, dem die Ausgangskonstellation 2 zugrunde gelegen hat, in der Endauswertung berücksichtigt. Für das Jahr 2011 hingegen, konnten Berechnungen gemacht

werden. Vergleichend mit der ersten Ausgangskonstellation, ergibt sich auch für die zweite Ausgangskonstellation eine mögliche Reduktion der Fallzahl bei der Wahl eines Signifikanzniveaus von 5% anstatt 2,5%. Bei Beibehaltung von Power und Hypothesenpaar ergibt sich eine Reduktion um 18,4%. Bei Betrachtung der Zahlen für das Jahr 2012 konnte unter den oben genannten Bedingungen eine Reduktion der Tierfallzahl von sogar 19,6% beobachtet werden. Für das Jahr 2013 kann bei einer Änderung des Signifikanzniveaus von 2,5% auf 5% ebenso eine Reduktion der Fallzahl um 22,6% festgestellt werden. Diese Reduktion der Fallzahl zeigt sich auch im Jahre 2014, wo ein mögliches Einsparungspotenzial von immerhin 1,43% der Tiere gezeigt werden konnte.

Ausgangskonstellation 3

Der Berechnung liegt folgende Ausgangskonstellation zu Grunde:

α = 5%  1-β = 80%  Hypothesenpaar = 1-seitig

Auch bei Betrachtung dieser Ausgangskonstellation kann bei Veränderung des Signifikanzniveaus von 5% auf 10% eine Reduktion der Tierfallzahl beobachtet werden. So ändert sich die Fallzahl im Jahr 2010 von 4.007 auf 3.593 Tiere, was einem prozentualen Anteil von 10,3% entspricht. Bei Betrachtung der Fallzahlen für das Jahr 2011 kann eine Verminderung der Tiere um 12,6% beobachtet werden. Analog dazu zeigt sich im Jahr 2012 eine mögliche Reduktion der Tierfallzahl um 18,6%. Bei einer Veränderung des Signifikanzniveaus hätten im Jahr 2014 sogar 26,2% der Tiere eingespart werden können. Für das Jahr 2013 bleibt zu erwähnen, dass kein entsprechendes Gutachten in der Endauswertung Berücksichtigung fand.

Ausgangskonstellation 4

Der Berechnung liegt folgende Ausgangskonstellation zu Grunde:

α = 1,67%  1-β = 80%  Hypothesenpaar = 2-seitig

Die Ausgangskonstellation 4 war, insgesamt betrachtet, eher selten, sodass sich in den Jahren 2010, 2013 und 2014 keines fand, das in der Endauswertung berücksichtigt wurde. Im Jahr 2011 hingegen konnte bei einer Erhöhung des Signifikanzniveaus von 1,67% auf 5% eine Reduktion der Tierfallzahl um 17,6% beobachtet werden. Analog dazu zeigt sich auch für das Jahr 2012 eine mögliche Verminderung der Fallzahl von in diesem Fall 22,2%.

# 4 Diskussion

In diesem Teil der Doktorarbeit werden die vorgestellten Ergebnisse mit den aktuellen Standards und geltenden Konventionen für Tierversuchsvorhaben in der medizinischen Grundlagenforschung, sowie mit bereits publizierten Texten zu diesem Thema verglichen und diskutiert.

## 4.1 Voraussetzungen

Die derzeit geltenden Konventionen für Tierversuche schlagen die Wahl eines Signifikanzniveaus von $\alpha = 5\%$ vor [25, 26, 44, 49, 52]. Die Werte für die Power sind etwas beliebiger wählbar, finden sich aber in der Regel zwischen 80 und 90, bzw. 95% [25, 26, 49, 52]. In den meisten Fällen liegen für Tierversuchsvorhaben zweiseitige Hypothesenpaare vor [25]. Gelegentlich finden sich aber auch gerichtete, einseitige Fragestellungen [25]. Dieses Vorherrschen der Ausgangskonstellation $\alpha = 5\%$, $1-\beta = 80\%$ bei einem zweiseitigen Hypothesenpaar spiegelt sich auch in den ausgewerteten Tierversuchsgutachten wider. So waren 74,1% der Gutachten dieser Ausgangskonstellation zuzuordnen. Für klinische Studien besagen die derzeit geltenden Richtlinien der EMA aus dem Jahre 1998, dass das Signifikanzniveau mit 5% und die Power mit 80% angenommen werden müssen und es sich im Rahmen des Versuchsdesigns um ein zweiseitiges Hypothesenpaar handeln soll [23]. Daraus lässt sich schließen, dass derzeit für Tierversuchsvorhaben dieselben strengen Richtlinien gelten wie für klinische Phase-III-Studien [49]. Anhand dieser Tatsache stellt sich die Frage, inwieweit solch strenge Richtlinien gerechtfertigt sind. Vor allem, weil Tierversuche in der Medizin ihren Platz hauptsächlich in der Grundlagenforschung haben. Die Grundlagenforschung an sich ist definiert als „zweckfreie, nicht auf unmittelbare praktische Anwendung hin betriebene Forschung, die sich mit den Grundlagen einer Wissenschaft o. Ä. beschäftigt“ [22]. Ihr Zweck ist somit die Hypothesen-findung für die weitere Forschung [20] und damit die „Gewinnung neuer Erkenntnisse, [die] die Basis für die angewandte Forschung bilden“ ([19], S. 9). Man kann ihr somit einen eher forschenden Charakter, als eine bestätigende Funktion zuschreiben [49]. Nicht außer Acht gelassen werden darf allerdings die Tatsache, dass die Ergebnisse, die mit Hilfe von Tierversuchen im Rahmen der Grundlagenforschung erzielt werden, als Grundlage für die weiteren Studienphasen der medizinischen Forschung dienen [19, 49]. Nicht zuletzt fordert

deshalb die für die Universität Ulm zuständige Tierschutzkommission, dass für Tierversuche im Rahmen der medizinischen Grundlagenforschung dieselben Standards gelten sollen, wie für Versuche am Menschen – vor allem mit dem Ziel, hoch valide Ergebnisse zu erhalten [42].

## 4.2 Die Rolle der statistischen Fallzahlplanung

Im Rahmen der Fragestellung dieser Doktorarbeit soll diskutiert werden, welche Rolle die statistische Fallzahlplanung für ebensolche Tierversuche der medizinischen Grundlagenforschung spielt. Welchen Teil kann sie dazu beitragen zu ermitteln, wie niedrig die tolerable Tierfallzahl sein kann, um trotzdem die dem Versuch zu Grunde liegende Fragestellung mit ausreichender Sicherheit beantworten zu können, um so eine solide Basis für die weitere Forschung zu bieten [44]. Immerhin stellt die formelle Fallzahlplanung nur eine Annäherung dar, die den Forschern erlaubt, von einer limitierten Menge an Informationen stabile Rückschlüsse auf die Grundgesamtheit zu ziehen und somit die Ergebnisse zu generalisieren [37].

Ihre grundlegende Berechtigung findet die formelle Fallzahlplanung zum einen in der Tatsache, dass es sich bei Tierversuchen der medizinischen Grundlagenforschung um sogenannte „induktive Studie[n]“ [[43], S. 369] handelt, d.h. um Studien, die dazu dienen, vom Einzelnen auf das Allgemeine zu schließen und die somit der Prüfung der Hypothesen dienen [43]. Zum anderen ist es auch aufgrund des öffentlichen Interesses an Tierversuchen wichtig, mittels einer validen biometrischen Fallzahlplanung nachweisen zu können, dass die benötigte Tierfallzahl durchaus ihre Berechtigung hat [43]. Des Weiteren gilt in Deutschland nach § 8 TierSchG die Regelung, dass Tierversuche im Rahmen der medizinischen Grundlagenforschung nur unter bestimmten Bedingungen durchgeführt und v.a. wiederholt werden dürfen [8]. Dazu muss überprüft werden, ob die bereits vorhandene Information, beispielweise aus anderen Tierversuchen, für die biometrische Fallzahlplanung des aktuellen Tierversuchsvorhabens verwendbar ist [42]. Um dies wahrscheinlicher zu machen sollte bei jedem Tierversuch, in Anlehnung an die ARRIVE-Guidelines, angegeben werden, wie die berechnete Tierfallzahl zustande gekommen ist und mit welcher statistischen Methode das Ergebnis ausgewertet wurde [40].

An dieser Stelle sollte auch erwähnt werden, dass die Tierfallzahl im Rahmen sogenannter Hauptversuche in der Forschung, im Gegensatz zur Fallzahl für Pilotstudien, nicht nach oben begrenzt ist. Vielmehr wird sie durch eine biometrische Berechnung, basierend auf den Angaben des Antragsstellers, abgeschätzt [42, 43]. Beachtung soll auch der Tatsache geschenkt werden, dass die optimale Tierfallzahl diejenige ist, die auf der Grundlage des für die Auswertung vorgesehenen statistischen Tests berechnet wurde [3]. Generell gilt, dass der Einsatz von zu wenigen Tieren dazu führen kann, dass ein vorhandener Effekt übersehen wird, allerdings die Verwendung von zu vielen Tieren zur Verschwendung derselben führt [24]. Demnach ist die statistische Fallzahlplanung die wahrscheinlich wichtigste Methode, um die Tierfallzahl passend zu berechnen und dadurch gegebenenfalls zu reduzieren [38]. Aus diesem Grund sollte es für die Veröffentlichung von Ergebnissen aus Tierversuchen allgemein gültige Richtlinien geben, die voraussetzen, dass genau dokumentiert wird, welche und wie viele Tiere verwendet werden. Genauso sollte einheitlich dokumentiert werden, wie die Tierfallzahl zustand kam und wie die Ergebnisse ausgewertet wurden [38, 40].

## 4.3 Die Möglichkeiten der statistischen Fallzahlplanung

Um die Tierfallzahl in der medizinischen Grundlagenforschung zu verringern stehen verschiedene biometrische Verfahren zur Verfügung, die im Folgenden vorgestellt werden.

Zunächst zur Betrachtung der statistischen Power und damit zur Betrachtung der Aussagekraft eines statistischen Tests. Verändert man die Power, so verändert sich gleichsinnig auch die Tierfallzahl. Eine Erniedrigung der Power führt zu einer Erniedrigung der Fallzahl. Das bedeutet im Umkehrschluss, dass eine größere Fallzahl generell zu einer größeren Power [26], damit eher zur Ablehnung einer falschen Alternativhypothese [15] und zu einer generell höheren Aussagekraft führt. Allerdings muss ein Studiendesign, welches im Rahmen von Tierversuchen dazu führt, dass aufgrund der großen Power selbst gewöhnliche, triviale Unterschiede entdeckt werden, im Hinblick auf die Tierfallzahl als verschwenderisch betrachtet werden [28]. Bei Betrachtung dieses Arguments darf aber nicht die Tatsache außer Acht gelassen werden, dass Studien mit einer zu geringen Power als Studien mit nur einem geringen Nutzen betrachtet werden.

Diese Aussage beruht auf der Annahme, dass Studien mit einer Chance von weniger als 80%, einen tatsächlich vorhandenen Unterschied zu entdecken als unethisch und damit nicht nutzbar gelten [1].

Alternativ kann man das zu testende Hypothesenpaar verändern. Generell ist es, in Bezug auf die Fallzahl, von Vorteil eine einseitige, gerichtete Testung durchzuführen, da in diesem Fall eine geringere Fallzahl für den Versuch verwendet werden muss [17]. Um allerdings ein solches Studiendesign zu ermöglichen sind diverse Vorinformationen von Nöten, die unter Umständen aus vorausgegangenen Tierversuchen, beispielsweise im Rahmen von Pilotstudien, generiert werden müssen. Für diese Pilotstudien ist, wie bereits in Kapitel 2.4 Biometrische Fallzahlplanung erwähnt, keine biometrische Fallzahlplanung möglich. Um einen inflationären Gebrauch von Tieren zu vermeiden wird deshalb die Fallzahl auf 60 bis 80 Tiere pro Pilotstudie begrenzt [43]. An dieser Stelle gilt es nochmals daran zu erinnern, dass ein Signifikanzniveau von 5% bei einem einseitigen Hypothesenpaar einem Signifikanzniveau von 10% bei zweiseitigem Hypothesenpaar entspricht. Dies begründet sich in der Tatsache, dass die gerichtete Forschung, ergo das einseitige Hypothesenpaar, die Richtung des Unterschieds testet und deshalb nur einen Ablehnungsbereich hat, der der Hälfte des Ablehnungsbereichs der zweiseitigen, ungerichteten Fragestellung entspricht [41].

Nicht zuletzt kann das Signifikanzniveau variiert werden. Hierbei verhält es sich so, dass eine Anhebung des Signifikanzniveaus, hin zu einem weniger strengen Grenzwert, zu einer Verminderung der Fallzahl führt [17]. Allerdings muss beachtet werden, dass diese Verminderung der Fallzahl auch zu einer geringeren Power führt [15]. Wird hingegen ein strengeres Signifikanzniveau gewählt, so kann dies mit einer Verminderung der Power gleichgesetzt werden, weshalb in einem solchen Fall die Fallzahl erhöht werden muss, um einen Powerverlust zu vermeiden [28, 44]. Bei Betrachtung der Tatsache, dass Tierversuche Bestandteil der Grundlagenforschung sind, scheint ein weniger strenges Signifikanzniveau von 10% für zweiseitige und 5% für einseitige Fragestellungen durchaus vertretbar [43]. Die Wahl eines 10%igen Signifikanzniveaus für zweiseitige Fragestellungen ist derzeit Standard in Studien der Pharmakokinetik, wenn es darum geht, die Frage der Bioäquivalenz zu klären [52, 55]. Diese Versuche sind, ebenso wie Tierversuche, der frühen Phase der klinischen Forschung zuzuordnen [43, 52]. Als

Gegenargument gilt es zu bedenken, dass ein zu großes Signifikanzniveau, beispielsweise ein α von 10%, zwar zu einer Erhöhung der Power und zu einer Erniedrigung der Tierfallzahl führen, aber dadurch gleichzeitig das Risiko für die Publikation eines Typ I-Fehlers steigt [28, 44]. Dies führe laut Fitts dazu, dass derselbe Versuch noch weitere Male durchgeführt werde und es so zu einer Verschwendung von Tieren komme [28]. Andererseits gibt er zu bedenken, dass die mehrmalige Durchführung desselben Experiments zu einem niedrigen Signifikanzniveau, beispielsweise einem α von 1%, ebenfalls zu einem hohen Verschleiß an Tieren führt [28]. Er kommt zu dem Schluss, dass, um das Gleichgewicht zwischen dem Fehler 1. Art und dem Fehler 2. Art zu wahren, das allgemein gültige Signifikanzniveau mit 5% gewählt werden sollte [28]. An dieser Stelle soll nochmals erwähnt werden, dass ein Fehler 1. Art, also ein falsch-positives Ergebnis, das einen nicht vorhandenen Unterschied bestätigt, schlimmere Folgen hat als ein Fehler 2. Art, da er direkte Auswirkungen auf die klinische Praxis hat [37]. Möchte man sowohl den Fehler 1. Art als auch den Fehler 2. Art so gering wie möglich halten, ist dies nur durch eine massive Erhöhung der Fallzahl möglich [28].

Um die Diskussion über die Wahl der Höhe des Signifikanzniveaus weiterzuführen, sei hier auf die Meinung von Bacchetti et al. verwiesen. Sie geben zu bedenken, dass „selbst bei Annahme der strittigen Voraussetzung, dass der Wert einer Studie allein von ihrer Power bestimmt wird […], sich die Balance zwischen dem Wert einer Studie und der für die Teilnehmer vertretbaren Belastung nicht verbessert“ [[1], S. 105]. Auch sind sie der Meinung, dass Studien auch dann, wenn sie kein Signifikanzniveau von 5% erreichen, als nützlich und wertvoll gelten können. Das sei dann der Fall, wenn ihre Ergebnisse wichtige Schätzwerte oder Konfidenzintervalle für weitere Studien liefern oder Eingang in Meta-Analysen finden [1].

Eine weitere, zu erwähnende, statistische Möglichkeit die Fallzahl für einen Versuch zu ermitteln, stellt die Berechnung der Fallzahl anhand von Konfidenzintervallen dar. Dies findet vor allem dann Anwendung, wenn die Genauigkeit, mit der ein Test die Zielgröße bestimmt, wichtiger ist, als das Ergebnis des Tests [32]. Soll die Zielgröße sehr genau bestimmt werden, so wird ein kleines Konfidenzintervall gewählt, was bedeutet, dass für diesen Test eine große Fallzahl von Nöten ist [32, 57]. Bei der Fallzahlbestimmung anhand von

Konfidenzintervallen wird jede Gruppe für sich betrachtet und kann mit Hilfe der entsprechenden Formel für das Konfidenzintervall bestimmt werden [32].

## 4.4 Das Versuchsdesign und die Verminderung der Tierfallzahl

Eine weitere biometrische Methode der Reduktion der Tierfallzahl, abgesehen von Veränderungen von Signifikanzniveau, Power und Hypothesenpaar, ist die Verwendung eines „*randomised block designs*“ [[26], S.7]. Dies ermöglicht eine bessere Kontrolle der Variabilität. Bei einem *randomised block design* oder *crossover design* sind dieselben Versuchseinheiten bzw. Tiere, Versuchs- und Kontrollgruppe. Das bedeutet, dass jede Versuchseinheit beispielsweise das Verum und das Placebo erhält, die Reihenfolge der Verabreichung aber zufällig bestimmt wird [24, 26]. Durch Anwendung eines *randomised block design* ist es möglich, bei weniger Varianz mehr Informationen zu erhalten, ohne dafür die Tierfallzahl zu erhöhen [27, 45]. Die während des Versuchs bestimmten Differenzen können dann statistisch ermittelt werden [26]. Aus diesem Grund sollte, wann immer möglich, ein solches Versuchsdesign verwendet werden [26]. Allerdings ist zu beachten, dass der mehrfache Einsatz derselben Tiere zwar insgesamt zu einer Verminderung der Fallzahl führt, aber sich dadurch gleichzeitig das Leid und die Schmerzen dieser Tiere erhöhen könnte [44].

Weitere Möglichkeiten die Tierfallzahl im Rahmen des Versuchsdesigns zu reduzieren bestehen darin, die interindividuelle Variabilität durch die Verwendung von Inzuchtstämmen zu vermindern. Des Weiteren sollten gesunde Tiere mit gleichem Gewicht und Alter verwendet werden [26].

Auch wenn für Tierversuche die geringstmögliche Fallzahl eingesetzt werden sollte, so muss trotzdem an einen „Sicherheitsfaktor“ gedacht werden. Auch sollten Ersatztiere für den Fall eines Ausfalls einzelner Tiere zur Gesamtfallzahl hinzugefügt werden [37].

Im Gegensatz dazu kann es manchmal durchaus sinnvoll sein, eine höhere Tierfallzahl zu verwenden. Dies ist dann der Fall, wenn durch den Einsatz einer höheren Fallzahl ein deutlich besseres Testergebnis in Bezug auf die Fragestellung erzielt werden kann. Findet ein solcher Test im Rahmen einer ganzen Testserie statt, so kann dadurch insgesamt eine Reduktion der Tierfallzahl erreicht werden, v.a. dann, wenn die Alternative darin besteht, viele kleine und schlechtere Versuche durchzuführen [44].

## 4.5 Die Aussagekraft kleiner Studien

In der Forschung hält sich die weit verbreitete Meinung, dass kleine Studien ethisch nicht vertretbar sind [1]. Dies beruht auf zwei nachvollziehbaren, aber unhaltbaren Argumenten [1]. Zum einen ist es richtig, dass größere Studien eine höhere Aussagekraft haben als kleine Studien [1]. Allerdings verteilt sich dann auch die Belastung im Rahmen der Studie auf eine größere Anzahl an Studienteilnehmern [1]. Eine Studie kann also nur als ethisch vertretbar betrachtet werden, wenn, betrachtet pro Teilnehmer, die Power größer ist, als die zu tragende Belastung [1]. Deshalb sollte für ethische Überlegungen die Aussagekraft pro Teilnehmer und nicht die Gesamtaussagekraft der Studie berücksichtigt werden [1]. Zum anderen wird in der Praxis eine Power von mindestens 80% mit der Gewissheit Erfolg zu haben, gleichgesetzt. Im Umkehrschluss wird eine Power von unter 80% mit der Gewissheit eines Misserfolgs gleichgesetzt [1]. Diese Zweiteilung erscheint im Rahmen einer ersten Schätzung durchaus sinnvoll, besonders dann, wenn das Signifikanzniveau auf 5% gesetzt wird und das Konfidenzintervall mit 95% angenommen wird. Allerdings ist sie nicht geeignet, um sich mit ethischen Aspekten oder der Fallzahl auseinander zu setzen [1]. So gilt, dass bei einem vorliegenden Studiendesign, bei dem eine Erhöhung der Fallzahl nicht zu einer Veränderung der Zusammensetzung innerhalb der Studiengruppen führt, beispielsweise im Rahmen einer Fall-Kontroll-Studie, eine Erhöhung der Fallzahl zu einer gleichsinnigen Erhöhung der Belastung für jeden einzelnen Teilnehmer führt [1]. Zusammenfassend kann also gesagt werden, dass größere Studien zwar die größere Power haben, aber gleichzeitig mehr Teilnehmer durch die Studie belasten [1]. Bacchetti et al argumentieren außerdem, dass bei steigender Teilnehmerzahl die Power pro Teilnehmer geringer wird, weshalb kleinere Fallzahlen eine bessere „ethische Balance“ aufweisen als Große ([1] , S. 106)

## 4.6 Das 3R-Prinzip und die Verminderung der Tierfallzahl

Auch wenn sich diese Arbeit primär mit den biometrischen Aspekten der Fallzahlplanung für Tierversuche in der medizinischen Grundlagenforschung auseinandersetzt, so soll doch das 3R-Prinzip nicht in Vergessenheit geraten, ist es doch das Grundprinzip, auf dem die Maßgabe zur statistischen Fallzahlplanung beruht [8, 16, 50]. So sollte die Tierfallzahl aufgrund ethischer, monetärer und

zeitlicher Gründe reduziert werden. Eine Möglichkeit, diese Reduktion zu erreichen, stellt die vorgestellte formelle Fallzahlplanung und spätere statistische Analyse dar [26]. Hierbei gestaltet sich folgendes Dilemma: Wie kann eine Verminderung der Fallzahl erreicht werden, ohne dass dies einerseits zu einer deutlichen Verminderung der Menge und Qualität der Informationen, die mittels des Experiments gewonnen werden kann, führt und andererseits die anderen zwei Prinzipien der 3R's – *Refinement* und *Replacement* – vernachlässigt werden [44]? Aus diesem Grund besteht die Notwendigkeit, ein Versuchsdesign und eine Analysemethode zu wählen, die am ehesten dazu geeignet sind, ein vertrauenswürdiges, valides und zur Fragestellung passendes Ergebnis zu liefern und dabei so effizient sind, dass nicht mehr Tiere als unbedingt nötig verwendet werden müssen [44]. Des Weiteren sollten statistische Möglichkeiten, wie beispielsweise Randomisierung und Verblindung angewendet werden, um eine Verzerrung der Ergebnisse zu vermeiden oder zu vermindern. Hierdurch kann ebenfalls eine Reduktion der Fallzahl erreicht werden, da potenzielle Störfaktoren schon vor Studienbeginn ausgeschaltet werden [44]. Dadurch wird das Versuchsergebnis valider und eine Wiederholung kann vermieden werden. Durch genaue Dokumentation kann die unnötige Wiederholung eines Experiments und damit der erneute Einsatz von Tieren vermieden werden, weshalb dies unbedingt zu beachten ist [44].

## 4.7 Ethische Aspekte und die Verminderung der Tierfallzahl

Auch ethische Aspekte dürfen bei der Diskussion über die Verminderung der Tierfallzahl nicht außer Acht gelassen werden. So ist doch der Erkenntnisgewinn in unserer Gesellschaft von hohem Stellenwert, weshalb Tierversuche unter bestimmten Voraussetzungen gerechtfertigt sind. Zu diesen Voraussetzungen zählt, neben beispielsweise der Unerlässlichkeit und dem Fehlen von Alternativen, der Einsatz von so wenigen Tieren wie möglich [19]. Trotzdem muss im Rahmen der Güterabwägung beachtet werden, dass Tierversuche nicht nur dem reinen Erkenntnisgewinn dienen. Vielmehr werden sie genutzt, um komplexe Vorgänge, wie beispielsweise die Funktionsweise der Sinnesorgane oder des Immunsystems zu verstehen, was letzten Endes nur unter Berücksichtigung des gesamten Organismus möglich ist [20]. Auch die Entwicklung wichtiger Therapien für Krankheiten, die derzeit unheilbar sind, ist ohne den Einsatz von Tierversuchen

derzeit nicht möglich [19]. Letztlich ist aus diesen Gründen aktuell kein kompletter Verzicht auf Tierversuche in der medizinischen Grundlagenforschung möglich [19].

## 4.8 Fazit

Die Wahl des Signifikanzniveaus α mit 5% erscheint nach Lektüre diverser Quellen als sehr fest in den Köpfen der Wissenschaftler und in den derzeit gültigen Standards verankert, sodass hier eine Veränderung schwierig erscheint. Zudem gilt es im Rahmen der Diskussion über eine Anhebung des Signifikanzniveaus von 5% auf 10% zwischen der Reduktion der Tierfallzahl und dem steigenden Risiko der Publikation eines Fehlers 1. Art, also der Veröffentlichung eines falsch-positiven Ergebnisses abzuwägen.

Auch eine Liberalisierung der Power auf 70% würde zu einer Reduktion der Fallzahl führen. Allerdings würde hierdurch die Aussagekraft der Studie abnehmen. Dies wiederum würde bedeuten, dass es auch in diesem Fall eher zur Annahme einer falschen Alternativhypothese und damit zur Publikation eines falsch-positiven Ergebnisses kommt. Generell wird in der heutigen Wissenschaft somit die Annahme der Power mit mindestens 80% gefordert [1, 45].

Aus diesen beiden Gründen soll abschließend als Kompromiss die gerichtete Testung, also die Testung zum einseitigen Hypothesenpaar vorgeschlagen werden. Diese Testung führt, ebenso wie eine Erhöhung des Signifikanzniveaus oder eine Erniedrigung der Power, zu einer Reduktion der Gesamtfallzahl, ohne dass die Auswertung an Aussagekraft verliert. Durch die gerichtete Forschung, die das Vorhandensein eines Unterschieds in einer vor der Versuchsdurchführung festgelegten Richtung untersucht, gibt es nur einen Ablehnungsbereich. Dies ermöglicht die Reduktion der Fallzahl [17]. Die Testung mittels einseitigem Hypothesenpaar wird auch von Herrn Professor Peter Pilz aus dem Fachbereich Biologie der Universität Tübingen in seiner Vorlesung „Statistische Grundlagen für die Versuchstierkunde“ vorgeschlagen [[45], Folie 15].

Diese Veränderung im Studiendesign führt, bei den im Rahmen dieser Doktorarbeit betrachteten Tierversuchsgutachten, zu folgendem Ergebnis: Bei Betrachtung der häufigsten und gängigsten Konstellation, der Ausgangskonstellation 1, zunächst aufsummiert für die Jahre 2010-2014, gelingt eine Reduktion der Fallzahl um ca. 13,9%, was einer Zahl von 6.262 Tieren entspricht. Diese Zahl von 6.262 Tieren hätte sich auf die einzelnen Jahre wie folgt

verteilt: Im Jahr 2010 hätte sich die Tierzahl auf 840 Tiere weniger belaufen, in 2011 wären 1.078 Tiere weniger zu verzeichnen gewesen. 2012 wären 1.405 Tiere weniger zum Einsatz gekommen, für 2013 wäre eine Reduktion um 236 Tiere möglich gewesen und im Jahre 2014 hätte eine Einsparung von 2.703 Tieren stattfinden können.

# 5 Zusammenfassung

Diese Dissertation behandelt das Thema „Biometrische Aspekte der Fallzahlplanung für Tierversuche in der medizinischen Grundlagenforschung“. Unter dem Begriff Tierversuch werden sämtliche Eingriffe oder Behandlungen subsummiert, die Schmerzen, Leid oder Schaden für die betroffenen Tiere bedeuten können. Als Basis für die Verankerung der biometrischen Fallzahlplanung in der Planung und Durchführung von Tierversuchen in der medizinischen Grundlagenforschung dient das 3R-Prinzip nach Russell und Burch von 1959. Es fordert die folgenden drei Dinge: *Replacement* - die Ersetzung von Tieren wann immer möglich, *Reduction* - die Reduktion der Fallzahl und *Refinement* - die Verfeinerung der Untersuchungs- und Versuchsmethoden. Es hebt somit die nicht zu unterschätzenden ethischen Aspekte der Tierversuche hervor. Die biometrische bzw. statistische Fallzahlplanung an sich findet ihre Begründung in verschieden Gesetzen, sowohl auf europäischer als auch auf Bundesebene.

Für die Berechnungen und Auswertungen dieser Arbeit liegen insgesamt 164 Tierversuchsgutachten mit einem Umfang von 76.620 Tieren im Zeitraum vom 01.01.2010 bis zum 31.12.2014 des Instituts für Epidemiologie und Medizinische Biometrie der Universität Ulm vor. Berücksichtigt wurden in der Endauswertung jedoch lediglich 112 Gutachten. All diesen Gutachten liegt eine statistische Fallzahlberechnung auf Grundlage der drei Parameter Signifikanzniveau $\alpha$, Power $1-\beta$ und Hypothesenpaar zugrunde. Diese drei Variablen wurden im Rahmen der Neuberechnungen verändert. Prinzipiell betrachtet ist das Optimum der Fallzahl als Kompromiss zwischen einer minimal ausreichenden und einer mindestens erforderlichen Fallzahl zu verstehen.

Die derzeit gültigen Konventionen der European Medicines Agency, kurz EMA, von 1998 verlangen ein Signifikanzniveau von 5%, eine Power von 80% und eine Testung zum zweiseitigen Hypothesenpaar. Dieser Vorgabe entsprachen unter allen berücksichtigten Gutachten 83 von 112 Gutachten. Um ein mögliches Einsparungspotenzial im Rahmen der Tierversuche in der medizinischen Grundlagenforschung aufzuzeigen, wurden verschiedene Neuberechnungen durchgeführt. Diese konnten zeigen, dass eine Anhebung des Signifikanzniveaus auf 10% zu einer Reduktion der Fallzahl um 13,9%, eine Liberalisierung der Power

auf 70% eine Reduktion um 11,3% und eine Veränderung des Versuchsdesigns hin zum einseitigen Hypothesenpaar ebenfalls eine Reduktion um 13,9% bewirkt. Die Veränderung des Versuchsdesigns hin zum einseitigen Hypothesenpaar ist das Fazit der Doktorarbeit und soll als möglicher Kompromiss zwischen der bestmöglichen Erfüllung des 3R-Prinzips (Replace, Reduce, Refine), der Bedingung der Verwendung der minimal möglichen und der mindestens erforderlichen Fallzahl und den derzeitig geltenden Konventionen mit Annahme eines 5%-igen Signifikanzniveaus und einer 80%-igen Power vorgeschlagen werden. Durch diese Maßnahme ist ein nicht unerhebliches Einsparungspotenzial realisierbar bei gleichzeitiger Beibehaltung der derzeit standardmäßig geltenden Werte für das Signifikanzniveau α und die Power 1-β.

# 6 Literaturverzeichnis

[1] Bacchetti P., Wolf L. E., Segal M. R., McCulloch C. E.: Ethics and Sample Size. American Journal of Epidemiology 161: S. 105-110 (2005).

[2] Balls M.: The Principles of Human Experimental Technique: Timeless Insights and Unheeded Warnings. Altex 27: S. 19-23 (2010).

[3] Barron A.: The Justification of Animal Numbers - the Role of Sample Size, Precision and Power Analysis in Assigning and Justifying Animal Numbers. Altex 28: S. 331-332 (2011).

[4] Bland M.: Chapter 9: Significance Tests. In: Bland M. (Hrsg.): An Introduction to Medical Statistics, 2. Aufl., Oxford University Press Inc., S. 133-151 (1995).

[5] Bland M. : Chapter 12: Methods Based on Rank Order. In: Bland M. (Hrsg.): An Introduction to Medical Statistics, 2. Aufl., Oxford University Press Inc., S. 205-224 (1995).

[6] Bland M.: Chapter 13: The Analysis of Cross-Tabulations. In: Bland M. (Hrsg.): An Introduction to Medical Statistics, 2. Aufl., Oxford University Press Inc., S. 225-253 (1995).

[7] Bundesministerium der Justiz und für Verbraucherschutz: Verordnung zum Schutz von zu Versuchszwecken oder zu anderen wissenschaftlichen Zwecken verwendeten Tieren (Tierschutz-Versuchstierverordnung - TierSchVersV). http://www.gesetze-im-internet.de/bundesrecht/tierschversv/gesamt.pdf (31.10.2015).

[8] Bundesministerium der Justiz und für Verbraucherschutz: Deutsches Tierschutzgesetz. http://www.gesetze-im-internet.de/tierschg/BJNR012770972.html (29.10.2005).

[9] Bundesministerium für Ernährung und Landwirtschaft: Versuchstierzahlen 2013. http://www.bmel.de/SharedDocs/Downloads/Broschueren/Tierschutzbericht-2015.pdf?__blob=publicationFile; (08.03.2016).

[10] Bundesministerium für Ernährung und Landwirtschaft: Versuchstierzahlen 2011. http://www.bmel.de/SharedDocs/Downloads/Broschueren/Tierschutzbericht-2015.pdf?__blob=publicationFile; (08.03.2016).

[11] Bundesministerium für Ernährung und Landwirtschaft: Versuchstierzahlen 2014. http://www.bmel.de/SharedDocs/Downloads/Broschueren/Tierschutzbericht-2015.pdf?__blob=publicationFile; (08.03.2016).

[12] Bundesministerium für Ernährung und Landwirtschaft: Versuchstierzahlen 2010. http://www.bmel.de/SharedDocs/Downloads/Broschueren/Tierschutzbericht-2015.pdf?__blob=publicationFile; (08.03.2016).

[13] Bundesministerium für Ernährung und Landwirtschaft: Tierschutzbericht der Bundesregierung 2015. http://www.bmel.de/SharedDocs/Downloads/Broschueren/Tierschutzbericht-2015.pdf?__blob=publicationFile; (08.03.2016).

[14] Bundesministerium für Ernährung und Landwirtschaft: Versuchstierzahlen 2012. http://www.bmel.de/SharedDocs/Downloads/Broschueren/Tierschutzbericht-2015.pdf?__blob=publicationFile; (08.03.2016).

[15] Cohen J.: Statistical Power Analysis. Current Directions in Psychological Science 1: S. 98-101 (1992).

[16] Das europäische Parlament und der Rat der europäischen Union: Richtlinie 2010/63/EU des Europäischen Parlaments und des Rates vom 22. September 2010 zum Schutz der für wissenschaftliche Zwecke verwendeten Tiere. http://www.bfr.bund.de/cm/343/5_Beratung_Anlage%203_2010-63-EU.pdf (29.10.2015).

[17] Dell R. B., Holleran S., Ramakrishnan R.: Sample Size Determination. ILAR Journal 43: S. 207-213 (2002).

[18] Der Landtag von Baden-Württemberg: Stellungnahme des Ministeriums für Ländlichen Raum und Verbraucherschutz: Tierversuche in Baden-Württemberg Drucksache 15/6077. http://www.landtag-bw.de/files/live/sites/LTBW/files/dokumente/WP15/Drucksachen/6000/15_6077_D.pdf (29.10.2015).

[19] Deutsche Forschungsgemeinschaft: Tierversuche in der Forschung. http://www.dfg.de/download/pdf/dfg_im_profil/geschaeftsstelle/publikationen/dfg_tierversuche_0300304.pdf (25.05.2016).

[20] Deutsche Forschungsgemeinschaft: Tierschutzbroschüre - "Tierversuche in der Forschung". http://www.dfg.de/download/pdf/dfg_magazin/forschungs

politik/tierschutz2015/tierversuchsbroschuere_doppelseitig.pdf (29.02.2016).

[21] Deutsches Primatenzentrum - Leibniz-Institut für Primatenforschung: Tierversuchszahlen in Deutschland. http://www.dpz.eu/de/abteilung/ueber-tierversuche/zahlen-und-fakten/tierversuchszahlen-in-deutschland.html; (08.03.2016).

[22] Duden: Grundlagenforschung, Die. http://www.duden.de/rechtschreibung /Grundlagenforschung (23.03.2016).

[23] European Medicines Agency. : ICH Topic E 9: Statistical Principles for Clinical Trials. http://www.ema.europa.eu/docs/en_GB/document_library/ Scientific_guideline/2009/09/WC500002928.pdf (22.03.2016).

[24] Festing M. F. W.: The Design and Statistical Analysis of Animal Experiments. ILAR Journal 43: S. 191-193 (2002).

[25] Festing M. F. W., Altman D. G.: Guidelines for the Design and Statistical Analysis of Experiments using Laboratory Animals. ILAR Journal 43: S. 244-258 (2002).

[26] Festing M. F. W.: How to Reduce the Number of Animals used in Research by Improving Experimental Design and Statistics. Anzccart Human Science Fact Sheet T10: S. 1-11 (2011).

[27] Festing M. F. W., Nevalainen T.: The Design and Statistical Analysis of Animal Experiments: Introduction to this Issue. ILAR Journal 55: S. 379-382 (2014).

[28] Fitts D. A.: Ethics and Animal Numbers: Informal Analyses, Uncertain Sample Sizes, Inefficient Replications, and Type I Errors. Journal of the American Association for Laboratory Animal Science 50: S. 445-453 (2011).

[29] Gaus W., Muche R.: Kapitel 2: Typen von Studien. In: Gaus W., Muche R. (Hrsg.): Medizinische Statistik - Angewandte Biometrie für Ärzte und Gesundheitsberufe, 1. Aufl., Schattauer GmbH, S. 37-56 (2014).

[30] Gaus W., Muche R.: Kapitel 14: Prinzip des Statistischen Tests. In: Gaus W., Muche R. (Hrsg.): Medizinische Statistik - Angewandte Biometrie für

Ärzte und Gesundheitsberufe, 1. Aufl., Schattauer GmbH, S. 264-280 (2014).

[31] Gaus W., Muche R.: Kapitel 18: T-Test. In: Gaus W., Muche R. (Hrsg.): Medizinische Statistik - Angewandte Biometrie für Ärzte und Gesundheitsberufe, 1. Aufl., Schattauer GmbH, S. 321-327 (2014).

[32] Gaus W., Muche R.: Kapitel 25: Fahlzahlbestimmung. In: Gaus W., Muche R. (Hrsg.): Medizinische Statistik - Angewandte Biometrie für Ärzte und Gesundheitsberufe, 1. Aufl., Schattauer GmbH, S. 444-463 (2014).

[33] Gaus W., Muche R.: Kapitel 17: Rangtests. In: Gaus W., Muche R. (Hrsg.): Medizinische Statistik - Angewandte Biometrie für Ärzte und Gesundheitsberufe, 1. Aufl., Schattauer GmbH, S. 302-320 (2014).

[34] Gaus W., Muche R.: Kapitel 15: Chi-Quadrat-Test und andere Tests für qualitative Zielgrößen. In: Gaus W., Muche R. (Hrsg.): Medizinische Statistik - Angewandte Biometrie für Ärzte und Gesundheitsberufe, 1. Aufl., Schattauer GmbH, S. 281-295 (2014).

[35] Gaus W., Muche R.: Kapitel 16: LogRank-Test. In: Gaus W., Muche R. (Hrsg.): Medizinische Statistik - Angewandte Biometrie für Ärzte und Gesundheitsberufe, 1. Aufl., Schattauer GmbH, S. 296-301 (2014).

[36] Gaus W., Muche R.: Kapitel 10: Überlebenszeitanalyse. In: Gaus W., Muche R. (Hrsg.): Medizinische Statistik - Angewandte Biometrie für Ärzte und Gesundheitsberufe, 1. Aufl., Schattauer GmbH, S. 188-209 (2014).

[37] Gogtay N. J.: Principles of Sample Size Calculation. Indian Journal of Ophthalmology 58: S. 517-518 (2010).

[38] Hess K. R.: Statistical Design Considerations in Animal Studies Published Recently in Cancer Research. Cancer Research 71: S. 625-625 (2011).

[39] Johnson P. D., Besselsen D. G.: Practical Aspects of Experimental Design in Animal Research. ILAR Journal 43: S. 202-206 (2002).

[40] Kilkenny C., Browne W., Cuthill I. C., Emerson M., Altman D. G.: Animal Research: Reporting in vivo Experiments: The ARRIVE Guidelines. British Journal of Pharmacology 160: S. 1577-1579 (2010).

[41] Lohringer H. : Grundlagen der Statistik - einseitige und zweiseitige Tests. http://www.statistics4u.com/fundstat_germ/cc_test_one_two_sided.html (03.03.2016).

[42] Mayer B., Muche R.: Chances and Obstacles of Pilot Projects in Animal Research for Statistical Sample Size Calculation. International Journal of Biological & Medical Research 3: S. 2319-2323 (2012).

[43] Mayer B., Muche R.: Die limitierte Aussagekraft formaler Fallzahlplanung im Rahmen von Tierversuchen der medizinischen Grundlagenforschung. Tierärztliche Praxis Kleintiere 6: S. 367-374 (2013).

[44] Parker R. M., Browne W. J.: The Place of Experimental Design and Statistics in the 3Rs. ILAR Journal 55: S. 477-485 (2014).

[45] Pilz P. : Statistische Grundlagen für die Versuchstierkunde. https://www.medizin.uni-tuebingen.de/tierschutz/Einfuhrung_in_die_Statstik_zur_Tierversuchsplanung_und_-auswertnug__Pilz_30.01.14__.pdf (07.01.2017).

[46] Rau M. : Aspekte und Methoden der Fallzahlplanung bei Tierversuchen. Bachelorarbeit, Hochschule Ulm (2011).

[47] Regierungspräsidium Tübingen. : Antrag auf Genehmigung bzw. Anzeige von Tierversuchen. https://rp.baden-wuerttemberg.de/Themen/Landwirtschaft/Documents/TierversucheAntrag.doc (07.01.2017).

[48] Regierungspräsidium Tübingen. : Angaben zur Biometrischen Planung. https://rp.baden-wuerttemberg.de/Themen/Landwirtschaft/Documents/BiometrischePlanung.doc (07.01.2017).

[49] Richter V., Mayer B., Muche R.: How Much Confidence do we need in Animal Experiments? Statistical Assumptions in Sample Size Estimation. Journal of Applied Animal Welfare Science, https://www.tandfonline.com/doi/abs/10.1080/10888705.2018.1423972?journalCode=haaw20 (09.07.2018)

[50] Russell W. M. S., Burch R. L.: The Principles of Humane Experimental Technique Chapter 4: THE SOURCES, INCIDENCE, AND REMOVAL OF INHUMANITY. http://altweb.jhsph.edu/pubs/books/humane_exp/chap4d (07.01.2017).

[51] Russell W. M. S., Burch R. L.: The Principles of Humane Experimental Technique. http://altweb.jhsph.edu/pubs/books/humane_exp/addendum (07.01.2017).

[52] Schuirmann D. J.: A Comparison of the Two One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability. Journal of Pharmacokinetics and Biopharmaceutics 15: S. 657-680 (1987).

[53] Statistical Solutions. : nQuery Adivsor + nTerim. http://www.statsols.com/products/nquery-advisor-nterim/ (07.03.2016).

[54] Tierforschungszentrum Ulm: Antragsformulare als Download. http://www.uni-ulm.de/einrichtungen/tfz/service/antragsformulare.html (16.04.2016).

[55] U.S. Department of Health and Human Services, Food and Drug Administration (FDA). : Guidance for Industry: Drug Interaction Studies - Study Design, Data Analysis, Implications for Dosing, and Labeling Recommendations (Draft Guidance). http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM292362.pdf; (16.03.2016).

[56] Weiß C.: Kapitel 9: Das Prinzip eines Statistischen Tests. In: Weiß C. (Hrsg.): Basiswissen Medizinische Statistik, 5. Aufl., Springer Medizin Verlag Heidelberg, S. 175-194 (2010).

[57] Weiß C.: Kapitel 8: Schätzverfahren. In: Weiß C. (Hrsg.): Basiswissen Medizinische Statistik, 5. Aufl., Springer Medizin Verlag Heidelberg, S. 157-174 (2010).

[58] Weiß C.: Kapitel 10: Lagetests. In: Weiß C. (Hrsg.): Basiswissen Medizinische Statistik, 5. Aufl., Springer Medizin Verlag Heidelberg, S. 195-238 (2010).

[59] Weiß C.: Kapitel 11: Tests zum Vergleich von Häufigkeiten. In: Weiß C. (Hrsg.): Basiswissen Medizinische Statistik, 5. Aufl., Springer Medizin Verlag Heidelberg, S. 217-238 (2010).

# 7 Abbildungs- und Tabellenverzeichnis

## 7.1 Abbildungsverzeichnis

Abbildung 1: Überblick über den Ablauf des Genehmigungsverfahrens zur Durchführung eines Tierversuchs [19, 39, 46] ..... 3

Abbildung 2: Entwicklung der Tierversuchszahlen in den Jahren 2010-2014 in Deutschland [9-14, 20, 21] ..... 4

Abbildung 3: Annahme - und Ablehnungsbereich beim t-Test; nach Vorlage von [41, 56] ..... 17

Abbildung 4: Startmenü nQuery Advisor® ..... 27

Abbildung 5: Maske für die Durchführung eines t-Tests mittels nQuery Advisor® 28

Abbildung 6: Maske für die Durchführung eines Wilcoxon-Tests mittels nQuery Advisor® ..... 28

Abbildung 7: Maske zur Berechnung der Wahrscheinlichkeit im Rahmen der Durchführung eines Wilcoxon-Tests ..... 29

Abbildung 8: Maske zur Berechnung der gemeinsamen Standardabweichung (Wilcoxon-Test) ..... 29

Abbildung 9: Maske zur Durchführung eines Chi-Quadrat-Tests mittels nQuery Advisor® ..... 30

Abbildung 10: Maske zur Durchführung eines LogRank-Tests mittels nQuery Advisor® ..... 30

Abbildung 11: Prozentuale Verteilung der betrachteten Tierversuchsgutachten auf die vier Ausgangskonstellationen ..... 34

## 7.2 Tabellenverzeichnis

Tabelle 1: Gesamtsumme der Tiere, aufgeschlüsselt nach Jahr und Gattung ..... 10

Tabelle 2: Anzahl der in der Endberechnung berücksichtigten Tiere, aufgeschlüsselt nach Jahr und Gattung ..... 11

Tabelle 3: Wahrheitstafel [56]..... 12

Tabelle 4: Darstellung der jeweils durchgeführten 18 Berechnungen und den entsprechenden Bedingungen..... 32

Tabelle 5: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 5%, 1-β = 80%, Hypothesenpaar = 2-seitig für die Jahre 2010-2014 (Tierversuchsgenehmigungen Universität Ulm)..... 37

Tabelle 6: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 2,5%, 1-β = 80%, Hypothesenpaar = 1-seitig für die Jahre 2010-2014 (Tierversuchsgenehmigungen Universität Ulm)..... 39

Tabelle 7: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 5%, 1-β = 80%, Hypothesenpaar = 1-seitig für die Jahre 2010-2014 (Tierversuchsgenehmigungen Universität Ulm)..... 40

Tabelle 8: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 1,67%, 1-β = 80%, Hypothesenpaar = 2-seitig für die Jahre 2010-2014 (Tierversuchsgenehmigungen Universität Ulm)..... 41

Tabelle 9: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 5%, 1-β = 80%, Hypothesenpaar = 2-seitig für das Jahr 2010 (Tierversuchsgenehmigungen Universität Ulm)..... 42

Tabelle 10: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 5%, 1-β = 80%, Hypothesenpaar = 2-seitig für das Jahr 2011 (Tierversuchsgenehmigungen Universität Ulm)..... 43

Tabelle 11: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 5%, 1-β = 80%, Hypothesenpaar = 2-seitig für das Jahr 2012 (Tierversuchsgenehmigungen Universität Ulm)..... 43

Tabelle 12: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 5%, 1-β = 80%, Hypothesenpaar = 2-seitig für das Jahr 2013 (Tierversuchsgenehmigungen Universität Ulm)..... 44

Tabelle 13: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 5%, 1-β = 80%, Hypothesenpaar = 2-seitig für das Jahr 2014 (Tierversuchsgenehmigungen Universität Ulm)..... 45

Tabelle 14: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 2,5%, 1-β = 80%, Hypothesenpaar = 1-seitig für das Jahr 2011 (Tierversuchsgenehmigungen Universität Ulm)..... 69

Tabelle 15: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 2,5%, 1-β = 80%, Hypothesenpaar = 1-seitig für das Jahr 2012 (Tierversuchsgenehmigungen Universität Ulm)..... 69

Tabelle 16: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 2,5%, 1-β = 80%, Hypothesenpaar = 1-seitig für das Jahr 2013 (Tierversuchsgenehmigungen Universität Ulm)..... 70

Tabelle 17: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 2,5%, 1-β = 80%, Hypothesenpaar = 1-seitig für das Jahr 2014 (Tierversuchsgenehmigungen Universität Ulm)..... 70

Tabelle 18: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 5%, 1-β = 80%, Hypothesenpaar = 1-seitig für das Jahr 2010 (Tierversuchsgenehmigungen Universität Ulm)..... 71

Tabelle 19: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 5%, 1-β = 80%, Hypothesenpaar = 1-seitig für das Jahr 2011 (Tierversuchsgenehmigungen Universität Ulm)..... 71

Tabelle 20: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 5%, 1-β = 80%, Hypothesenpaar = 1-seitig für das Jahr 2012 (Tierversuchsgenehmigungen Universität Ulm)..... 72

Tabelle 21: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 5%, 1-β = 80%, Hypothesenpaar = 1-seitig für das Jahr 2014 (Tierversuchsgenehmigungen Universität Ulm)..... 72

Tabelle 22: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 1,67%, 1-β = 80%, Hypothesenpaar = 2-seitig für das Jahr 2011 (Tierversuchsgenehmigungen Universität Ulm)..... 73

Tabelle 23: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 1,67%, 1-β = 80%, Hypothesenpaar = 2-seitig für das Jahr 2012 (Tierversuchsgenehmigungen Universität Ulm)..... 73

# 8 Anhang

## 8.1 Vergleich der Fallzahlen – sortiert nach Ausgangskonstellation und Jahr

### *8.1.1 Ausgangskonstellation 2*

α = 2,5%  1-β = 80%  Hypothesenpaar = 1-seitig

Jahr 2011

*Tabelle 14: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 2,5%, 1-β = 80%, Hypothesenpaar = 1-seitig für das Jahr 2011 (Tierversuchsgenehmigungen Universität Ulm)*

| | | Power 1-β in % | | | | | |
|---|---|---|---|---|---|---|---|
| | | 70 | 80 | 90 | 70 | 80 | 90 |
| Signifikanz-niveau α in % | 1 | 2.575 (+3,4) | 2.911 (+16,9) | 3.475 (+39,6) | 3.044 (+22,3) | 3.469 (+39,3) | 3.905 (+56,8) |
| | 5 | 1.809 (-27,3) | 2.032 (-18,4) | 2.350 (-5,6) | 2.134 (-14,3) | **2.490** | 2.790 (+12,1) |
| | 10 | 1.588 (-36,2) | 1.652 (-33,7) | 1.955 (-21,5) | 1.809 (-27,3) | 2.032 (-18,4) | 2.350 (-5,6) |
| | | 1-seitig | | | 2-seitig | | |
| | | Hypothesenpaar | | | | | |

Jahr 2012

*Tabelle 15: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 2,5%, 1-β = 80%, Hypothesenpaar = 1-seitig für das Jahr 2012 (Tierversuchsgenehmigungen Universität Ulm)*

| | | Power 1-β in % | | | | | |
|---|---|---|---|---|---|---|---|
| | | 70 | 80 | 90 | 70 | 80 | 90 |
| Signifikanz-niveau α in % | 1 | 1.087 (+9,8) | 1.281 (+29,4) | 1.572 (+58,8) | 1.281 (+29,4) | 1.475 (+49,0) | 1.863 (+88,2) |
| | 5 | 602 (-39,2) | 796 (-19,6) | 990 (±0) | 796 (-19,6) | **990** | 1.281 (+29,4) |
| | 10 | 505 (-49,0) | 602 (-39,2) | 796 (-19,6) | 602 (-39,2) | 796 (-19,6) | 990 (±0) |
| | | 1-seitig | | | 2-seitig | | |
| | | Hypothesenpaar | | | | | |

## Jahr 2013

*Tabelle 16: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 2,5%, 1-β = 80%, Hypothesenpaar = 1-seitig für das Jahr 2013 (Tierversuchsgenehmigungen Universität Ulm)*

| | | Power 1-β in % | | | | | |
|---|---|---|---|---|---|---|---|
| | | 70 | 80 | 90 | 70 | 80 | 90 |
| Signifikanz-niveau α in % | 1 | 1.021 (+16,4) | 1.054 (+20,2) | 1.252 (+42,8) | 1.054 (+20,2) | 1.246 (+42,1) | 1.310 (+49,4) |
| | 5 | 673 (-23,3) | 679 (-22,6) | 744 (-15,2) | 833 (-5,0) | **877** | 1.048 (+19,5) |
| | 10 | 635 (-27,6) | 641 (-26,9) | 679 (-22,6) | 673 (-23,3) | 679 (-22,6) | 877 (±0) |
| | | 1-seitig | | | 2-seitig | | |
| | | Hypothesenpaar | | | | | |

## Jahr 2014

*Tabelle 17: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 2,5%, 1-β = 80%, Hypothesenpaar = 1-seitig für das Jahr 2014 (Tierversuchsgenehmigungen Universität Ulm)*

| | | Power 1-β in % | | | | | |
|---|---|---|---|---|---|---|---|
| | | 70 | 80 | 90 | 70 | 80 | 90 |
| Signifikanz-niveau α in % | 1 | 2.650 (+18,7) | 2.666 (+19,4) | 3.131 (+40,3) | 3.083 (+38,1) | 3.115 (+39,6) | 3.565 (+59,7) |
| | 5 | 1.766 (-20,9) | 2.200 (-1,4) | 2.232 (±0) | 2.200 (-1,4) | **2.232** | 2.666 (+19,4) |
| | 10 | 1.333 (-40,3) | 1.750 (-21,6) | 1.782 (-20,2) | 1.766 (-20,9) | 2.200 (-1,4) | 2.232 (±0) |
| | | 1-seitig | | | 2-seitig | | |
| | | Hypothesenpaar | | | | | |

### *8.1.2 Ausgangskonstellation 3*

α = 5%  1-β = 80%  Hypothesenpaar = 1-seitig

Jahr 2010

*Tabelle 18: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 5%, 1-β = 80%, Hypothesenpaar = 1-seitig für das Jahr 2010 (Tierversuchsgenehmigungen Universität Ulm)*

| | | Power 1-β in % | | | | | |
|---|---|---|---|---|---|---|---|
| | | 70 | 80 | 90 | 70 | 80 | 90 |
| Signifikanz-niveau α in % | 1 | 4.705 (+17,4) | 5.349 (+33,5) | 6.117 (+52,7) | 5.306 (+32,4) | 5.845 (+45,9) | 6.867 (+71,4) |
| | 5 | 3.732 (-6,9) | **4.007** | 4.705 (+17,4) | 4.115 (+2,7) | 4.634 (+15,7) | 5.388 (+34,5) |
| | 10 | 3.407 (-15,0) | 3.593 (-10,3) | 4.552 (+13,6) | 3.732 (-6,9) | 4.007 (±0) | 4.705 (+17,4) |
| | | 1-seitig | | | 2-seitig | | |
| | | Hypothesenpaar | | | | | |

Jahr 2011

*Tabelle 19: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 5%, 1-β = 80%, Hypothesenpaar = 1-seitig für das Jahr 2011 (Tierversuchsgenehmigungen Universität Ulm)*

| | | Power 1-β in % | | | | | |
|---|---|---|---|---|---|---|---|
| | | 70 | 80 | 90 | 70 | 80 | 90 |
| Signifikanz-niveau α in % | 1 | 1.703 (+20,8) | 1.949 (+38,2) | 2.252 (+59,7) | 2.003 (+42,1) | 2.066 (+46,5) | 2.423 (+71,8) |
| | 5 | 1.347 (-4,5) | **1.410** | 1.712 (+21,4) | 1.523 (+8,0) | 1.589 (+12,7) | 1.889 (+34,0) |
| | 10 | 1.170 (-17,0) | 1.233 (-12,6) | 1.473 (+4,5) | 1.347 (-4,5) | 1.410 () | 1.652 (+17,2) |
| | | 1-seitig | | | 2-seitig | | |
| | | Hypothesenpaar | | | | | |

## Jahr 2012

*Tabelle 20: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 5%, 1-β = 80%, Hypothesenpaar = 1-seitig für das Jahr 2012 (Tierversuchsgenehmigungen Universität Ulm)*

| | | Power 1-β in % | | | | | |
|---|---|---|---|---|---|---|---|
| | | 70 | 80 | 90 | 70 | 80 | 90 |
| Signifikanz-niveau α in % | 1 | 1.743 (+32,8) | 2.084 (+58,7) | 2.435 (+85,5) | 2.064 (+57,2) | 2.226 (+69,5) | 2.614 (+99,1) |
| | 5 | 1.249 (-4,9) | **1.313** | 1.657 (+26,2) | 1.344 (+2,4) | 1.721 (+31,1) | 1.994 (+51,9) |
| | 10 | 1.012 (-22,9) | 1.069 (-18,6) | 1.321 (+0,6) | 1.249 (-4,9) | 1.313 (±0) | 1.657 (+26,2) |
| | | 1-seitig | | | 2-seitig | | |
| | | Hypothesenpaar | | | | | |

## Jahr 2014

*Tabelle 21: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 5%, 1-β = 80%, Hypothesenpaar = 1-seitig für das Jahr 2014 (Tierversuchsgenehmigungen Universität Ulm)*

| | | Power 1-β in % | | | | | |
|---|---|---|---|---|---|---|---|
| | | 70 | 80 | 90 | 70 | 80 | 90 |
| Signifikanz-niveau α in % | 1 | 2.098 (+31,0) | 2.385 (+49,0) | 2.930 (+83,0) | 2.357 (+47,2) | 2.714 (+69,5) | 3.252 (+103,1) |
| | 5 | 1.314 (-17,9) | **1.601** | 1.958 (+22,3) | 1.609 (+0,5) | 1.895 (+18,4) | 2.281 (+42,5) |
| | 10 | 1.056 (-34,0) | 1.182 (-26,2) | 1.469 (-8,2) | 1.314 (-17,9) | 1.601 (±0) | 1.958 (+22,3) |
| | | 1-seitig | | | 2-seitig | | |
| | | Hypothesenpaar | | | | | |

### *8.1.3 Ausgangskonstellation 4*

α = 1,67%  1-β = 80%  Hypothesenpaar = 2-seitig

Jahr 2011

*Tabelle 22: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 1,67%, 1-β = 80%, Hypothesenpaar = 2-seitig für das Jahr 2011 (Tierversuchsgenehmigungen Universität Ulm)*

| | | Power 1-β in % | | | | | |
|---|---|---|---|---|---|---|---|
| | | 70 | 80 | 90 | 70 | 80 | 90 |
| Signifikanz-niveau α in % | 1,67 | 28 (-17,6) | 31 (-8,8) | 34 (±0) | 31 (-8,8) | **34** | 40 (+17,7) |
| | 5 | 22 (-35,3) | 25 (-26,5) | 28 (-17,6) | 25 (-26,5) | 28 (-17,6) | 34 (±0) |
| | 10 | 19 (-44,1) | 22 (-35,3) | 25 (-26,5) | 22 (-35,3) | 25 (-26,5) | 28 (-17,6) |
| | | 1-seitig | | | 2-seitig | | |
| | | Hypothesenpaar | | | | | |

Jahr 2012

*Tabelle 23: Ergebnisse der Berechnungen als Fallzahlen und Änderung dieser im Vergleich zur Ausgangsfallzahl in Prozent (Angabe in Klammern) für die Ausgangskonstellation α = 1,67%, 1-β = 80%, Hypothesenpaar = 2-seitig für das Jahr 2012 (Tierversuchsgenehmigungen Universität Ulm)*

| | | Power 1-β in % | | | | | |
|---|---|---|---|---|---|---|---|
| | | 70 | 80 | 90 | 70 | 80 | 90 |
| Signifikanz-niveau α in % | 1,67 | 84 (-22,2) | 90 (-16,7) | 108 (±0) | 90 (-16,7) | **108** | 126 (+16,7) |
| | 5 | 72 (-33,3) | 78 (-27,8) | 84 (-22,2) | 78 (-27,8) | 84 (-22,2) | 102 (-5,6) |
| | 10 | 54 (-50,0) | 60 (-44,4) | 78 (-27,8) | 54 (-50,0) | 78 (-27,8) | 84 (-22,2) |
| | | 1-seitig | | | 2-seitig | | |
| | | Hypothesenpaar | | | | | |

Teile dieser Dissertation wurden bereits in folgenden Fachartikeln veröffentlicht:

Veronika Richter, Rainer Muche & Benjamin Mayer: How much confidence do we need in animal experiments? Statistical assumptions in sample size estimation. Journal of Applied Animal Welfare Science, ISSN: 1088-8705 (Print) 1532-7604 (Online) Journal homepage: https://www.tandfonline.com/doi/abs/10.1080/10888705.2018.1423972?journalCode=haaw20 (2018)

## Lebenslauf

Der Lebenslauf wurde aus Gründen des Datenschutzes entfernt