UNIVERSITÉ DE GENÈVE

Archive ouverte UNIGE
https://archive-ouverte.unige.ch

Thèse de privat-docent 2025

Open Access

This version of the publication is provided by the author(s) and made available in accordance with the copyright holder(s).

Advances in management of frequent endocrine tumors in clinical practice : pituitary adenomas, thyroid nodules

Mavromati, Maria

How to cite

MAVROMATI, Maria. Advances in management of frequent endocrine tumors in clinical practice : pituitary adenomas, thyroid nodules. Thèse de privat-docent, 2025. doi: 10.13097/archive-ouverte/unige:190375

This publication URL: https://archive-ouverte.unige.ch/unige:190375
Publication DOI: 10.13097/archive-ouverte/unige:190375

© This document is protected by copyright. Please refer to copyright holder(s) for terms of use.

UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE

Clinical Medicine Section
Department of Medicine

"Advances in management of frequent endocrine tumors in clinical practice: pituitary adenomas, thyroid nodules "

Thesis submitted to the Faculty of Medicine of
the University of Geneva

for the degree of Privat-Docent
by

Maria MAVROMATI

Geneva

2025

Table of Contents

1. ABBREVIATIONS

ACR-TIRADS: American College of Radiology Thyroid Imaging Reporting and Data System
ACTH: adrenocorticotropin
AI: artificial intelligence
ATA: American Thyroid Association
AVP: arginine vasopressin
BMI: body mass index
CI: confidence intervals
CT: computed tomography
DA: dopamine agonists
EU-TIRADS: European Thyroid Imaging Reporting and Data System
FNA: fine-needle aspiration
FSH: follicle-stimulating hormone
FDG PET: 18F-fluorodeoxyglucose positron emission tomography
GH: growth hormone
GHRH: growth hormone releasing hormone
HR: hazard ratio
IGF-1: insulin-like growth factor 1
IGF-BP: insulin-like growth factor binding protein
K-TIRADS: Korean Thyroid Imaging Reporting and Data System
LC-MS: liquid chromatography mass spectrometry
LH: luteinizing hormone
LLN: lower limit of normal
MEN: multiple endocrine neoplasia
MRI: magnetic resonance imaging
NF: non-functioning
NFPA: non-functioning pituitary adenoma
OGTT: oral glucose tolerance test
PET-CT: positron emission tomography computed tomography
Pit-NET: pituitary neuroendocrine tumor
ROM: rate of malignancy
ROSE: rapid on-site evaluation

RSS: risk stratification system

SD: standard deviation

SDS: standard deviation score

SMR: standardized mortality ratio

SRL: somatostatin receptor ligand

TIRADS: Thyroid Imaging Reporting and Data System

TSH: thyroid stimulating hormone

ULN: upper limit of normal

US: ultrasound

WHO: World Health Organization

2. ACKNOWLEDGMENTS

I would like to express my deepest gratitude to Prof. Sophie Leboulleux for being a source of inspiration, motivation and continuous support, that were fundamental for the realization of this work.

I am also grateful to

Prof. Jacques Philippe, Prof. Philippe Chanson, Prof. Alain Golay, Prof. François Jornayvaz for their valuable guidance

Prof. Shahan Momjian, Prof. Frédéric Triponez, Dr Marco Demarchi, Dre Essia Saiji, Dr Claudio de Vito, Dr Eugenio Fernandez as well as all the medical and nursing staff of my department, for their precious collaboration

Dr Patrick Meyer and Prof. Agathocles Tsatsoulis who inspired me to study endocrinology

The Fond National Suisse (FNS) and Fond de Département des Spécialités de Médecine (DMIG/DSM) of Geneva University Hospital for the funding

My parents for their ongoing support and perpetual encouragement

Finally, this work could not have been accomplished without the love and unfailing support of my family, my husband Manos and our son Yiannis. This memoire is dedicated to them.

3. SUMMARY

Pituitary adenomas are benign tumors that are diagnosed incidentally during imaging performed for other reasons, or due to symptoms related to local compression, hormonal hypersecretion or hypopituitarism. While surgery is the treatment of choice for growth-hormone (GH) – secreting, adrenocorticotropin (ACTH) – secreting and thyrotropin (TSH) - secreting pituitary adenomas, the main indication for surgery of a clinically non-functioning adenoma (NFPA) is visual impairment or threat. However, hypopituitarism in patients with NFPA is increasingly recognized as a relative indication for surgery.

Growth hormone disorders (acromegaly and GH deficiency) are assessed using growth hormone (GH) and Insulin Growth factor (IGF-1) levels measurement. Therefore, IGF-1 reference normative values are essential for diagnosis, treatment efficacy evaluation and follow-up.

Thyroid nodules are found in up to 50% of the adult general population but only 5 to 10% are malignant. Their management aims at recognizing malignancy, based on ultrasound risk stratification scores and molecular tests of thyroid cytology. However, diagnostic procedure has limitations, resulting in unnecessary fine-needle aspiration (FNA) cytology and unnecessary surgery in a considerable number of patients.

4. GENERAL CONSIDERATIONS:

Frequent endocrine tumors in clinical practice (pituitary adenomas, thyroid nodules)

Pituitary adenomas, or pituitary neuroendocrine tumors (pit-NETs) are more frequent than they were believed to be in the past and are often discovered as incidentalomas due to the wider use of imaging. While surgery is first-line treatment for clinically functioning pit-NETs, except for prolactinomas which are treated medically, surgery is indicated for non-functioning pituitary adenomas (NFPAs) mostly in case of visual defects due to compression of the optic nerves or imminent visual threat. Hypopituitarism because of a non-functioning (NF) macroadenoma has traditionally been considered as only a relative indication for surgery. However, the benefits of surgery for pituitary function in the absence of visual impairment are currently recognized based on clinical cohorts showing frequent recovery of hypopituitarism with surgery. Furthermore, diagnosis and follow-up of patients with growth hormone (GH) disorders (acromegaly, GH-deficiency) by means of insulin-like growth factor 1 (IGF-1) levels are often complex, due to the variability of IGF-1 measurement among different commercial immunoassays. Thus, several studies have focused on the standardization of IGF-1 assays and the elaboration of age- and sex-specific reference normative data for IGF-1 with easier calculation of standard deviation (SD) scores.

Thyroid nodules are the most frequent endocrine tumor in clinical practice with a prevalence of > 50% in imaging series, and with a 5-13% malignancy rate when discovered incidentally. Ultrasound (US) risk stratification scores (RSS) help stratify malignancy risk, based on the presence of high-risk features but have low reproducibility and specificity, leading to unnecessary fine-needle aspiration (FNA) procedures. FNA cytology specimens can also be non-diagnostic (Bethesda III or IV) and despite their low malignancy rates, current guidelines suggest diagnostic thyroid lobectomy. Refinement of US RSSs, quantitative evaluation of unnecessary surgery rates as well as real-life cost-effectiveness studies for molecular testing of thyroid cytology are the first steps towards improvement in the management of patients with thyroid nodules.

5. PITUITARY ADENOMAS

a. Introduction

i. *The pituitary gland*

The pituitary gland (hypophysis), located in the sella turcica of the cranial base, is a small endocrine organ responsible to produce hormones that control growth, energy balance and metabolism, reproduction, response to stress, lactation, as well as water and sodium balance. It is divided into the anterior lobe (adenohypophysis) and the posterior lobe (neurohypophysis). The anterior lobe produces adrenocorticotropin (ACTH), which stimulates the adrenal cortex for cortisol and androgen production, thyroid-stimulating hormone (TSH), which stimulates the thyroid gland for thyroid hormone production, growth hormone (GH), responsible for growth and different aspects of metabolism, gonadotropins (follicle-stimulating hormone – FSH and luteinizing hormone – LH), which stimulate gonads, and prolactin, which is essential for lactation (Figure 1). The posterior lobe serves as storage for antidiuretic hormone or arginine vasopressin (AVP) and oxytocin, produced by the hypothalamus. The pituitary gland communicates with the hypothalamus, receiving stimulating and inhibitory signals through the pituitary stalk. Neighboring anatomic structures are the optic chiasm, and sphenoid sinuses, comprising cranial nerves (Figure 2) [1].

Adenomas are the most frequent affection of the pituitary gland and are of benign nature, while infiltrative disease or tumors with more aggressive potential are rare[2,3]. Pituitary adenomas are characterized as microadenomas if their larger diameter is less than 10 mm and macroadenomas if their larger diameter is at least 10 mm in size. In addition, they can be non-functioning or cause clinically significant symptoms in case of hormonal hypersecretion. Thus, pituitary adenomas can be totally asymptomatic if they are of small size and non-functioning[4]. On the contrary, larger adenomas can cause symptoms by impairing pituitary hormonal production (hypopituitarism) through compression of normal pituitary cells or of the pituitary stalk, as well as symptoms related to the compression of adjacent structures such as the optic chiasm and nerves (visual field defects or visual acuity impairment), the cavernous sinus (oculomotor nerve palsies) and meninges (headaches)[5]. Clinically functioning pituitary adenomas, whether they are microadenomas or macroadenomas, induce symptoms related to the excessive production of pituitary hormones (acromegaly, Cushing's disease etc.). Diagnosis of pituitary adenomas can thus

be incidental (incidentalomas), can be made because of symptoms of hormonal deficiency or excess, or can be driven by compression symptoms, mostly visual problems[4,5]. Finally, macroadenomas can sometimes present with symptoms of apoplexy, due to hemorrhage or infraction, including sudden headaches, visual defects or meningeal irritation.

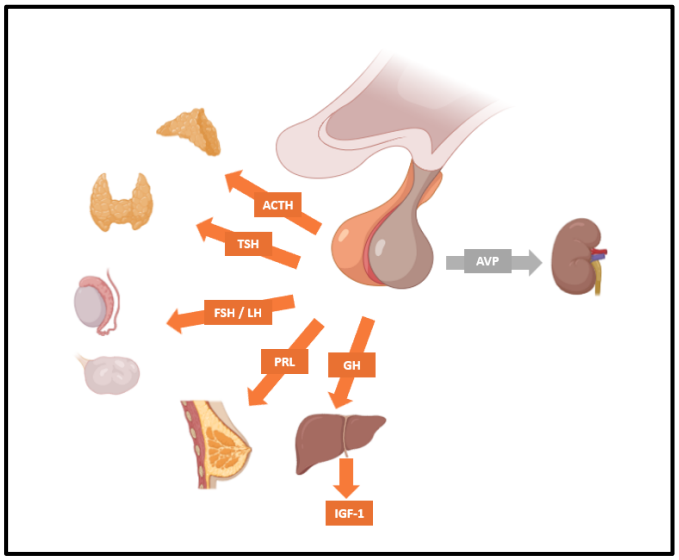

Figure 1: Main hormones produced by the anterior (ACTH, TSH, FSH and LH, PRL, GH) and the posterior (AVP) pituitary and their target organs (Created in BioRender)

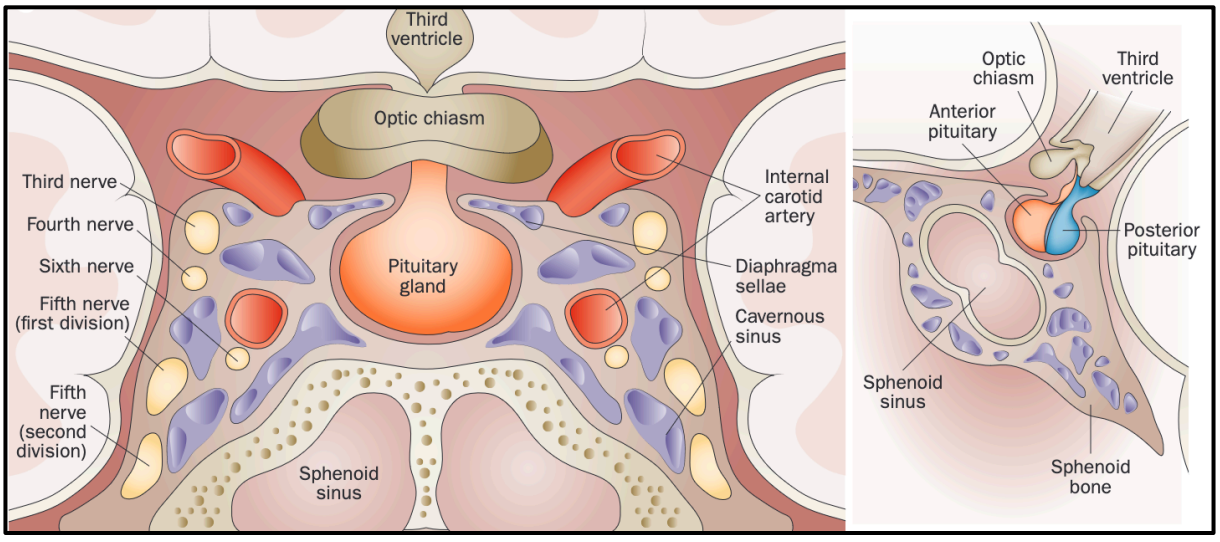

Figure 2: Normal anatomy of the sellar and parasellar regions surrounding the pituitary gland, coronal (left) and sagittal views (right) (Reproduced from Di Ieva A. et al. Nat Rev Endocrinol 2014;10(7);423-35 [1])

ii. *Subtypes and epidemiology of pituitary adenomas*

The term pituitary neuro-endocrine tumors (pit-NET) has recently been suggested as more precise than the term pituitary adenomas [6].

Upon diagnosis, classification among different subtypes is based on symptoms and signs related to hormonal excess. Hormonal workup in blood and urine confirms differential diagnosis between clinically functioning and non-functioning pituitary adenomas. Pit-NETs are thus classified as prolactinomas (excess prolactin production which should be differentiated from mild hyperprolactinemia due to stalk compression), clinically non-functioning adenomas, somatotroph adenomas (GH secretion causing acromegaly), corticotroph adenomas (ACTH secretion, Cushing's disease), and rarely, clinically functioning gonadotroph and mixed adenomas[6].

However, histological examination of the tumor can detect adenomas with positive immunohistochemical staining to pituitary hormones even if there is no proof of clinically relevant hormonal secretion in the circulation and no symptoms related to hormonal excess; this qualifies a pit-NET as clinically silent. In addition, the wider use of transcription factors in immunohistochemistry evaluation, mainly PIT-1, TPIT, SF1, has led to a better classification of pit-NETs according to cell lineage [6,7]. Pit-NETs are thus classified as PIT-1-lineage pit-NETs (among which somatotrophs, lactotrophs and thyrotrophs), TPIT-lineage pit-NETs (corticotrophs), SF1-lineage pit-NETs (gonadotrophs), and pit-NETs without distinct cell lineage (pluri-hormonal and null cell) [6].

The prevalence of pituitary adenomas in autopsy series reaches 10%, still, most of them are very small and clinically irrelevant and 39% are non-functioning. Imaging series find a higher prevalence of about 22.5% but the majority are also microadenomas [8-10]. Recent populational studies find higher prevalence rates than older series. A British registry found a prevalence of 78 pituitary adenomas per 100 000 inhabitants among which 57% were prolactinomas, 28% non-functioning, 11% growth-hormone producing adenomas, 2% corticotropin producing adenomas and 2% of unknown functional status [11]. Non-functioning pituitary adenomas were the most common subtype in men (57%) and prolactinomas, the most common subtype in women (76%) in this registry. In patients younger than 60 years old, prolactinomas were the most prevalent subtype (60%) while older patients had mostly non-functioning tumors (61%) [11]. Another registry from Iceland, including files between 1955 and 2012, found a higher prevalence rate of 116 cases per 100 000 inhabitants and somewhat similar distribution among subtypes, with prolactinomas being the most prevalent in the general population and mostly among women and younger patients while non-functioning adenomas were the most frequent in men and older individuals [12]. Interestingly, this study found increasing incidence rates with

time and mostly of non-functioning adenomas and authors argue that this cannot be attributed only to the wider use of imaging techniques [12].

iii. *Treatment of functioning pitNETs: a brief overview*

Except for prolactinomas, which are managed medically with dopamine agonists, first line treatment for functioning pit-NETs is surgery [13].

GH-secreting adenomas

Acromegaly is mainly caused by growth hormone secreting pituitary adenomas (somatotroph or GH-secreting adenomas) while ectopic production of GH or growth-hormone releasing hormone (GHRH) is extremely rare [14]. Recent populational studies have shown that prevalence of acromegaly is higher than it was previously believed to be while a German registry has found a prevalence as high as 1 case per 1000 inhabitants [11,15-17]. GH exerts its anabolic effects on muscle and bone to stimulate growth and its metabolic effects on adipose tissue (lipolysis) and the liver (gluconeogenesis). Anabolic effects on muscle and bone are mostly mediated by IGF-1, which is either produced by GH in the liver (circulating IGF-1) or in the target tissues (paracrine effect), while GH metabolic effects on adipose tissue and the liver are mainly direct [14,18]. Due to GH and IGF-1 excess, patients with acromegaly suffer from acral enlargement and coarse features, sweating and headaches, carpal tunnel syndrome, arthritis, impaired glucose tolerance or diabetes, hypertension, obstructive sleep apnea, colon polyps and cardiologic complications such as valve disease and cardiomyopathy [18,19]. Acromegaly is thus associated with decreased survival and causes of death in older series are mainly cardiovascular and respiratory complications, while hypertension and diabetes also seem to participate in decreased survival rates [20,21]. Treatment of acromegaly improves prognosis reaching that of healthy general population as GH levels approach normal range; good biologic control of the disease is thus necessary [22,23]. Recent studies, such as the French Acromegaly Registry, have demonstrated improved disease control over time [24]. A Swedish Registry showed a decrease in mortality over time, and mortality was mainly due to circulatory disease and malignancy [25]. Recent studies also show increased cancer risk in acromegaly patients, and this is possibly related to improved management and prolonged survival over time [26].

Transsphenoidal surgery is the first line treatment in acromegaly [18]. In case of failure to achieve biologic control with surgery alone, medical treatment is indicated [18]. However, surgery is useful even if complete tumor resection is improbable, since surgical debulking is associated to higher rates of biologic control with medical treatment [27]. Medical treatment in acromegaly is performed mainly with somatostatin receptor ligands (first generation:

octreotide, lanreotide, or second generation: pasireotide) or the GH-receptor antagonist pegvisomant [18,28]. Dopamine agonists, which are first-line treatment for prolactinomas can also be used in selected cases, if GH excess is mild [18,28]. Somatostatin receptor ligands (SRL) and dopamine agonists have central actions on the level of the pituitary adenoma causing direct inhibition of GH secretion and thus, indirect inhibition of IGF-1 production in the liver [29]. First generation SRL octreotide and lanreotide offer biologic control of acromegaly in 40-50% of patients as well as >20% reduction in adenoma volume in > 50% of patients [30-32]. Still, 50% of patients do not reach biochemical control [33]. Second generation SRL pasireotide offers biochemical control of 20% of patients who were not controlled with first generation SRL, as well as >20% reduction in adenoma volume in 80% of patients but often results in hyperglycemia and diabetes that can sometimes be difficult to control [34,35]. SRL are administered in the form of monthly injections, but oral octreotide has recently been approved by the FDA and has been shown to maintain biochemical control in patients previously controlled by injectable SRL [36]. GH receptor antagonist pegvisomant, which is administered in the form of daily subcutaneous injections, acts by blocking the GH receptor and directly inhibits IGF-1 secretion [29]. It has thus no action on adenoma volume but offers biochemical control in 65-80% of patients [37,38]. Combined treatment of SRL and pegvisomant seems to be the most effective in achieving biochemical control in acromegaly [39]. Finally, radiotherapy, either fractionated or stereotactic radiosurgery, can be considered if medical treatment is ineffective or not tolerated and while biochemical control is latent, it is achieved in 75% of patients after 10 years [40].

ACTH-secreting adenomas

ACTH-secreting pituitary adenomas (corticotroph adenomas, Cushing's disease) represent the most frequent cause of endogenous Cushing's syndrome (60-70%), followed by Cushing's syndrome of adrenal origin (20-30%) while ectopic causes are rare (5-10%) [41]. Diagnosis is made in most cases because of symptoms and signs of cortisol excess, such as facial plethora, purple wide striae, skin atrophy and easy bruising, central obesity, supra-clavicular fat pads, 'buffalo hump', proximal muscle wasting, or because of complications related to cortisol excess including thrombo-embolic disease, osteoporosis and pathologic fractures, frequent infections, peptic ulcer, neuro-psychiatric morbidity such as depression and anxious disorders, as well as metabolic complications (diabetes, hypertension and cardio-vascular morbidity). Cushing's disease is a rare disease with an estimated incidence of 1.2-2.4 cases / million / year, while it is more frequent in women than in men with a pic at 25-45 years of age [42,43]. Cushing's syndrome is associated with increased mortality, and hazard ratio for all-cause mortality in Cushing's disease is 2.3

(95% CI 1.7-3.0), mostly due to infections, cardio-vascular and thrombo-embolic events [44]. In addition, biologic cure decreases mortality rates without normalizing them and excess mortality seems to persist in patients with remission (HR: 1.9, 95% CI 1.5-2.3) [45].

First line treatment of Cushing's disease is surgery with selective transsphenoidal adenomectomy by an experienced neurosurgeon. However, complete resection of the corticotroph adenoma is not always easy as ACTH-producing adenomas as frequently very small in size while pituitary MRI is normal in 25-40% of cases despite biologic confirmation of a pituitary cause of Cushing's syndrome. Thus, identification of the adenoma often requires surgical exploration of the pituitary gland and para-sellar region. Remission rates after surgery are high (60-80%), still, 5-20% of patients will relapse, among which 50% in the first 5 years after surgery [46,47]. In these cases, a second surgery can be considered if there is a visible and non-invasive residual or relapsing adenoma. However, medical treatment is often required.

Medical treatment for Cushing's disease is performed with pituitary targeting agents, such as the second generation SRL pasireotide or with adrenal targeting agents. Among adrenal targeting agents, the oral steroidogenesis inhibitor osilodrostat, which has been approved and widely commercialized in 2020, has shown the best efficacy rates with more than 75% of patients normalizing mean 24h-urine free cortisol in clinical trials [48,49]. Other older adrenal targeting molecules include ketoconazole and metyrapone, but their use is currently decreasing in clinical practice due to lower efficacy and longer half-life. Patients require long-term follow-up to ensure control of hypercortisolism and related complications and to detect and treat potential relapse. Pituitary radiotherapy, either fractionated or stereotactic radiosurgery can be considered in inoperable or recurrent Cushing's disease with a 50-83% remission rate achieved between 6 months and 3 years, but with high risk of hypopituitarism [50]. Finally, bilateral adrenalectomy can be an option in refractory cases of severe, life-threatening hypercortisolism.

Prolactinomas

Prolactin secreting adenomas are the only pituitary adenomas not requiring surgery as first line treatment. They respond well to dopamine agonists (DA) cabergoline, bromocriptine and quinagolide. Cabergoline is the most widely used dopamine agonist, due to its longer half-life that does not require daily administration, as well as to its increase efficacy rates compared to the other DA [51].

Clinical manifestations of prolactinomas include symptoms related to prolactin excess such as hypogonadism, infertility and galactorrhoea but also compressive mass effects (visual defects, hypopituitarism, headaches) in patients with large tumors. Goals of treatment with DA in patients with prolactinomas thus include normalization of prolactin

levels which will restore gonadal function and stop galactorrhoea in women but also tumor shrinkage [52].

Treatment with oral DA has high efficacy rates and prolactin normalizes in 76% of patients treated with bromocriptine and 89% of patients treated with cabergoline [53]. Other studies have reported prolactin normalization and tumor shrinkage with cabergoline in 80% of patients with macroprolactinomas and 95% of patients with microprolactinomas [53]. Still, resistance to DA, defined as failure to achieve normal prolactin levels and ≥ 50% of tumor size decrease is reported in 10-20% of patients treated with cabergoline [54-56]. Surgery is indicated in these cases as well as in patients with DA intolerance, such as nausea and vomiting, headaches, dizziness or compulsive behaviour, even though side effects are rare with cabergoline [51,52].

iv. *Treatment of non-functioning pitNETs: effect of surgery on pituitary function*

Surgery for non-functioning pituitary adenomas is indicated in case of visual defects due to compression of the optic nerves or chiasm but also in case of visual threat if imaging shows proximity or direct contact with the tumor [13]. Whether the presence of hypopituitarism should be an indication for surgery or not has been controversial. Still, recent studies, have shown that surgery performed for non-functioning macroadenomas by experienced neurosurgeons more frequently improves pituitary function than deteriorates it. To examine the potential benefit of surgery in patients with NFPAs and no visual impairment one should evaluate the growth potential of these tumors, complications of surgery, impact of hypopituitarism, whether due to tumor itself or postoperative, as well as the effect of surgery on pituitary function in different studies. Finally, prognostic factors of new pituitary dysfunction and improvement, by hormonal axis, are important to guide decision making.

Growth rates in NF-pitNETs that did not have surgery generally depend on reasons of diagnosis (incidental or not), clinical presentation (asymptomatic or symptomatic) as well as their size upon diagnosis (microadenoma or macroadenoma) [57-59]. Microadenomas rarely grow during follow-up and only 5% of them become macroadenomas, while growth velocity is estimated at 0.4 mm/year [60]. On the other hand, macroadenomas have a higher growth potential and 20-60% of them will grow during a 2-7 years follow-up [61]. Growth velocity for macroadenomas is estimated at 1 mm / year and those in contact with the optic chiasm will more frequently grow (73% vs 29%) [61-63]. Still, some studies have shown a spontaneous decrease in volume of NF macroadenomas, which could be partially attributed to subclinical apoplectic events [61,62]. Imaging characteristics as well as proliferation markers (Ki67, p53, mitosis count) in histological examination after surgery

have also been associated with growth potential [64]. For those NF macroadenomas that have not been subject to surgery, risk for new pituitary dysfunction is estimated at 12% per year and risk of apoplexy at 1% per year (10% throughout follow-up) [61,62,65,66].

Serious complications of transsphenoidal surgery for NFPAs are rare but non-negligible. Mortality rates are ≤ 1% while other serious complications such as meningitis, cerebrospinal fluid leakage, persistent arginine vasopressin deficiency (previously named central diabetes insipidus) and visual impairment appear in ≤ 5% of cases [67]. Surgical transsphenoidal techniques, endoscopy and miscroscopy, seem to be equivalent in terms of complication rates and postoperative pituitary function but there has been no prospective direct comparison between them [68-71]. In addition, postoperative results, regarding size of residual tumor, seem to be superior if surgery for a growing NFPA is performed earlier, before the appearance of symptoms [72,73].

Patients with NFPAs have increased morbidity and mortality rates and this is due either to the adenoma itself or to surgical treatment. Results from a Swedish registry including 2975 patients with NFPAs, among which 52% had surgery, have shown a small but significant increase in mortality with a standardized mortality ratio (SMR) or 1.10 (95% CI: 1.00-1.20), which was more important in patients younger than 40 years of age (SMR: 2.68, 95% CI: 1.23-5.09) and in women (SMR: 1.29, 95% CI: 1.1-1.48) [74]. Causes of death were mostly cerebrovascular and infectious diseases [74]. A British registry including 546 patients who had surgery for NFPA and with a median follow-up of 8 years showed increased mortality with SMR of 3.5 (95% CI: 2.8-4.4) while causes of death were cardiovascular, infectious and oncologic [75]. In this registry, patients who were diagnosed after the age of 50 had increased mortality (hazard ratio: 1.1, 95% CI 1.07-1.13, $p<0.001$) and there was no impact of hypopituitarism on mortality [75]. Hypopituitarism per se also has an impact on mortality rates. Treatment of central adrenal insufficiency with supraphysiologic doses of glucocorticoids has been associated with increased mortality as was untreated growth hormone deficiency and insufficient treatment of central hypothyroidism [76-79]. Mortality also seems to increase with the number of deficient hormonal axis, being higher in patients with panhypopituitarism [78]. In summary, data show higher mortality rates in patients with NFPAs, whether they had surgery or not, while the most systematically accounted risk factors are female gender and treatment of central adrenal insufficiency with supraphysiologic doses of glucocorticoids [79].

Studies evaluating the impact of transsphenoidal surgery on pituitary function have examined the risk of new pituitary insufficiencies, the probability of recovery from preoperative hormonal deficits, as well as risk factors for worsening or improvement of

pituitary function with surgery. However, most studies have performed a more global evaluation and separate data by hormonal axis are scarce.

Two small studies from Spain and the USA, published before 2000, have found 35% and 33% of recovery of at least one axis after surgery respectively [80,81]. The Spanish study has shown that the ACTH axis was the most fragile to the negative effects of surgery, while the American study showed that 22% of patients had new hormonal deficiencies after surgery with risk factors being larger adenoma size and preoperative hypopituitarism [80,81]. In 2004, a German cohort including 660 patients, showed the best results with only 4% new postoperative deficiencies, and more than 40% recovery, which was more frequent in patients with tumors smaller than 3 cm and preoperative hyperprolactinemia [82]. High recovery rates with surgery were also shown in a study from the US including 160 patients with NFPAs, among which 55% had improvement or complete normalization of pituitary function which could occur several months after surgery (median time of recovery was 12.2 months) [83].

The fact that improvements of pituitary function can occur 6-12 months after surgery in some patients has also been shown in a French and Belgian cohort including 246 patients with NFPAs, among which 80% had hypopituitarism at baseline [84]. The percentage of patients with hypopituitarism dropped to 69% at 3 months and 61% at 1 year. Fifty percent of patients with preoperative hypopituitarism had improved pituitary function in 1 year. Hyperprolactinemia at diagnosis and smaller tumors were predictive of better outcomes in terms of pituitary function in this cohort [84].

Recent cohorts provide more information of new deficiencies and recovery by hormonal axis but have contrasting results. In two studies from the USA including 209 and 305 patients who had transsphenoidal surgery for NFPAs, new ACTH-deficiency was found to be the rarest event (9.7%) in the first study and the most frequent event (6%) in the second study [85,86]. Furthermore, ACTH-deficiency had the higher recovery rate (44.3%) in the first study but the lowest (3%) in the second [85,86].

A study from Geneva University Hospital including 137 patients who had surgery for NFPAs, showed a 25% improvement and 7% worsening of global pituitary function postoperatively and provided detailed information about the effects of surgery on each different hormonal axis separately, with clear biologic definition of deficiencies and stratification of patients who did not have proper screening or dynamic testing [87]. In this cohort, the ACTH axis was found to be the most fragile, with 9.6% new deficiencies, while the TSH axis was found to be the most resistant with only 1.6% new postoperative deficiencies. New hormonal deficiency of at least one axis was found in 10% of patients and recovery of at least one axis was shown in 46% of patients with growth hormone

deficiency and central adrenal insufficiency being the most and least probable to recover. High prolactin levels before surgery were associated with higher recovery rates [87].

The finding that preoperative hyperprolactinemia can predict recovery of pituitary dysfunction with surgery has been consistent in several studies, among which the French and Belgian cohort and the Geneva cohort [84,87]. A possible explanation could be the fact that patients with preoperative hyperprolactinemia have hypopituitarism due to compression of the pituitary stalk, rather than destruction of anterior pituitary cells, with resolution of the compression effect after surgery [88].

The largest cohort reported so far was a Swedish registry including 838 patients who were operated for a NFPA [89]. One year after surgery, 23% had new ACTH-deficiency while 26% of patients with preoperative ACTH-deficiency recovered. Recoveries in the TSH and FSH / LH axis occurred in 14% and 15% of patients while new deficiencies for these axes were shown in 22 and 29% of patients respectively [89]. Contrary to previous studies, the Swedish registry showed somewhat lower recovery rates than new deficiencies. Still, this is probably related to the nature of the cohort that did not include only expert surgeons.

Taken together, studies on the effect of transsphenoidal surgery for NFPAs show low pituitary function worsening rates, while several patients experience recovery of at least one axis [90]. Preoperative hyperprolactinemia seems to be a marker predicting improvement of pituitary function with surgery. Thus, hypopituitarism can be considered as a relative indication for surgery in patients with non-functioning pituitary macroadenomas and no visual impairment or threat.

v. *Diagnosis and follow-up of growth-hormone disorders (acromegaly, growth-hormone deficiency) by means of IGF-1 measurement*

IGF-1 measurement has been traditionally used for the diagnosis and follow-up of acromegaly, in association with GH (random or after oral glucose tolerance test – OGTT). The Endocrine Society 2014 Guidelines suggest using GH levels after OGTT and IGF-1 levels for diagnosis as well as IGF-1 levels and random GH as a therapeutic efficacy target [18]. However, there is a recent shift towards relying mainly on IGF-1 levels, with the 2024 Consensus Statement suggesting the use of GH measurements to confirm diagnosis only if IGF-1 results are equivocal, while follow-up relies exclusively on IGF-1 monitoring [91]. For growth-hormone (GH) deficiency, the 2016 Endocrine Society Guidelines for the treatment of hypopituitarism and the more recent American Association of Clinical Endocrinologists and American College of Endocrinology guidelines for GH deficiency diagnosis and treatment, recommend establishing diagnosis of adult GH deficiency based on low IGF-1

levels in patients with organic disease and at least 3 pituitary hormonal deficiencies or the presence of genetic defects known to affect the GH axis, without the need of GH dynamic testing, while follow-up on GH treatment relies mainly on IGF-1 values [92,93].

However, there is considerable variability among different commercial methods for IGF-1 measurement, which impacts patient classification to normal, low or high IGF-1 levels, and influences diagnosis of GH disorders (acromegaly, GH-deficiency) and follow-up of treatment efficacy [94]. Variability is observed among different assays but also on the same assay performed by different laboratories [95]. Such variability is explained by pre-analytical conditions, analytical conditions related to the characteristics of each assay, but also normative reference values suggested by each assay and laboratory as well as the population from which they have been derived.

One source of analytical variability of IGF-1 assays is related to the type of calibrator used and the 2011 WHO consensus statement on standardization and evaluation of GH and IGF-1 assays suggests standardization against the same calibrator, 02/254, a recombinant international IGF-1 standard preparation [96,97]. Other reasons explaining analytical variability include antibodies sensitivity and specificity as well as differences among methods used to remove IGF-1 binding proteins that can interfere with the measurement [98,99].

Still, even if analytical variability is minimized, for example even if the same international calibrator is used, IGF-1 results differ among immunoassays. Several studies show that this is mainly related to the fact that reference normative values are obtained from different populations [100]. The WHO consensus statement of IGF-1 assays standardization recommends obtaining normative values from a healthy reference population with representation of all age groups [97]. However, differences in the inclusion criteria of individuals among different populations result in differences among obtained normative values. Indeed, IGF-1 levels vary with age and sex, increasing significantly during puberty, especially in men, then declining quickly during the 2nd and 3rd decade, then more slowly thereafter. IGF-1 levels also vary with BMI, hormonal treatments, diabetes, and renal and hepatic function, and thus different normal values are obtained depending on the way the reference population has been chosen. In addition, the distribution of IGF-1 raw values in a given population is not Gaussian, and for calculation of standard deviation scores (SDSs), it is necessary to perform a transformation to a normal, Gaussian, distribution.

Conforming to the 2011 Consensus statement for GH and IGF-1 assay standardization, an international multicenter study, obtained and published, in 2014, age- and sex-specific reference IGF-1 values for iSYS (Immunodiagnostic Systems), an automated immunoassay [97,101]. The study included 15 041 healthy subjects from the United States, Europe and Canada, from birth to 94 years of age, and showed a progressive increase of

IGF-1 during childhood with a peak around 15 years, then a progressive decline, as well as slightly lower IGF-1 values in women [101].

In 2016, the VARIETE cohort aimed to establish normal reference IGF-1 values for 6 widely used commercial immunoassays for the adult French general population [102]. Following the 2011 recommendations, 911 healthy adults, with representation of all age groups, were included, in a cross-sectional manner, from 10 French centers [102]. Subjects had a detailed medical history and clinical examination and were enrolled after exclusion of medications and conditions that influence IGF-1 levels. The cohort finally included approximately 100 subjects per age decade, and sex- and age- specific normal values were calculated. Since the distribution of IGF-1 values is not Gaussian, age- and sex-specific curves were normalized with Cox-Box power transformation and normative reference values ranging from percentile 2.5 to percentile 97.5 were obtained. A calculator for SDS was also issued. Still, even though they were obtained from the same healthy reference population, concordance of reference IGF-1 values among methods was moderate to good. Women were found to have higher IGF-1 values than men, with all 6 immunoassays, until the age of 50, then lower, which is possibly due to the exclusion of individuals receiving oral estrogens known to reduce IGF-1 levels. Finally, there was poor concordance between reference intervals obtained by the study and those proposed by each kit manufacturer, which highlights the importance of defining reference normative values from a well-defined large population, including all age groups [102].

The application of normative reference values obtained with the VARIETE study for the classification of patients with GH disorders was further tested in a cohort of 102 patients (56 patients with acromegaly, 14 patients with GH deficiency and 32 patients with suspected GH disorder) [103]. In each patient, IGF-1 values were measured with all 6 assay kits that were included in the VARIETE study and pairwise concordance between the assays were calculated both for IGF-1 raw data and SD scores. Still, even though normative data were obtained from the same large, healthy, well calibrated population, concordance between assays remained variable, especially for IGF-1 levels close to the normal range. These findings support the fact that patient follow-up is ideally performed with the same assay, or at least with assays sharing the same analytical characteristics[103].

Differences in IGF-1 normative data have also been observed between different ethnicities. Indeed, a large population-based study including more than 1.4 million individuals from Europe and the United States (USA) that had IGF-1 measurements with the iSYS assay has found significant differences in age-specific normative reference values, with the US population having higher upper limit of normal (ULN) levels by 15-20% compared to European population [104]. This finding further complexifies the definition of reference values with wider utility.

Variability in patient classification to low, normal or high IGF-1 categories is not a major issue upon diagnosis of patients with acromegaly (when symptoms and signs are helpful for confirmation) but becomes a significant clinical problem after surgery, as it complexifies the identification of patients who remain incurred and would benefit from a medical treatment as well as for treatment titration during follow-up. In patients treated with somatostatin receptor ligands (SRLs), IGF-1 is higher just before the next injection and variability among IGF-1 values is higher in patients with uncontrolled acromegaly [105].

IGF-1 can also be measured with liquid chromatography tandem mass spectrometry (LC-MS) methods, which give normative values that resemble those obtained with immunoassays [106-108]. Nevertheless, LC-MS methods are more expensive and more complex to calibrate than immunoassays and variability is not significantly improved compared with immunoassays [107,108].

In conclusion, difficulties in recognizing limitations of IGF-1 measurement assays can lead to inappropriate therapeutic decisions. In order to minimize variability, laboratories should prefer immunoassays calibrated against the WHO 02/254 international standard and carefully choose adequate methods for binding protein elimination. Clinicians must be familiar with the specific characteristics of the assays used in their practice and implement age- and sex-specific reference normative values derived from large healthy populations, also taking into account the variability of IGF-1 levels in an individual patient. Therapeutic decisions in patients with slightly increased IGF-1 levels whether for treatment initiation or titration, must also consider the patient's clinical picture and the impact of the above changes on quality of life.

b. Publications

i. The impact of transsphenoidal surgery on pituitary function in patients with non-functioning macroadenomas

Non-functioning pituitary adenomas (NFPAs) are the second most frequent subtype of pituitary adenomas after prolactinomas in the general population, and the most common subtype in men and older adults. While NFPAs with a size < 10 mm (microadenomas) will rarely grow and require management, those with a size ≥ 10 mm (macroadenomas) tend to progress and cause hormonal insufficiencies from pituitary dysfunction, as well as visual impairment from optic chiasm compression. According to guidelines, transsphenoidal surgery for NF macroadenomas is indicated mainly in case of visual impairment or threat and preoperative hypopituitarism is only considered a relative indication. Surgery itself can be the cause of new postoperative hormonal insufficiencies and specific predictors of improvement or worsening of pituitary function have not been identified.

The objective of this study was to describe the impact of transsphenoidal surgery on pituitary function in patients with NF macroadenomas and search for predictors of postoperative recovery of hormonal production or new deficiencies, by axis.

We reviewed files from 310 consecutive transsphenoidal surgeries performed from March 2004 to January 2018 from the same experienced surgeon and included 137 patients with NF macroadenomas with median tumor size of 24.8 mm. Before surgery, 58.4% of patients had visual impairment and 67% had at least one hormonal axis defect with central hypogonadism being the most frequent. After surgery, 46% of patients showed recovery of at least one hormonal axis with growth hormone deficiency and central adrenal insufficiency being the most and least probable to recover, respectively (45.5% and 15.4%). Only 10% of patients had new deficiency in at least one hormonal axis, with ACTH production being the most fragile (9.2%) and TSH production the most resistant (1.6%) to the effects of surgery. Men and patients with high prolactin levels preoperatively were more likely to recover from hypopituitarism with surgery while no prognostic factors for new hormonal deficiencies were found.

This real-life cohort of patients who had transsphenoidal surgery for NF macroadenoma in a tertiary center shows that postoperative recovery of pituitary function is more frequent than the appearance of new hormonal deficiencies, thus, pituitary dysfunction can be considered a relative indication for surgery.

Endocrine (2023) 81:340–348
https://doi.org/10.1007/s12020-023-03400-z

ORIGINAL ARTICLE

The impact of transsphenoidal surgery on pituitary function in patients with non-functioning macroadenomas

Maria Mavromati[1] · Thomas Mavrakanas[2] · François R. Jornayvaz[1] · Karl Schaller[3] · Aikaterini Fitsiori[4] · Maria I. Vargas[4] · Johannes A. Lobrinus[5] · Doron Merkler[5] · Kristof Egervari[5] · Jacques Philippe[6] · Sophie Leboulleux[1] · Shahan Momjian[3]

Received: 1 February 2023 / Accepted: 10 May 2023 / Published online: 24 May 2023
© The Author(s) 2023

Abstract

Purpose Transsphenoidal surgery for non-functioning pituitary adenomas (NFPAs) can alter pituitary function. We assessed the rates of improvement and deterioration of pituitary function by axis and searched for predictive factors of these outcomes.

Methods We reviewed consecutive medical files from patients having had transsphenoidal surgery for NFPA between 2004 and 2018. Pituitary functions and MRI imaging were analyzed prior and after surgery. The occurrence of recovery and new deficit were documented per axis. Prognostic factors of hormonal recovery and new deficits were searched.

Results Among 137 patients analyzed, median tumor size of the NFPA was 24.8 mm and 58.4% of patients presented visual impairment. Before surgery, 91 patients (67%) had at least one abnormal pituitary axis (hypogonadism: 62.4%; hypothyroidism: 41%, adrenal insufficiency: 30.8%, growth hormone deficiency: 29.9%; increased prolactin: 50.8%). Following surgery, the recovery rate of pituitary deficiency of one axis or more was 46% and the rate of new pituitary deficiency was 10%. Rates of LH-FSH, TSH, ACTH and GH deficiency recovery were 35.7%, 30.4%, 15.4%, and 45.5% respectively. Rates of new LH-FSH, TSH, ACTH and GH deficiencies were 8.3%, 1.6%, 9.2% and 5.1% respectively. Altogether, 24.6% of patients had a global pituitary function improvement and only 7% had pituitary function worsening after surgery. Male patients and patients with hyperprolactinemia upon diagnosis were more likely to experience pituitary function recovery. No prognostic factors for the risk of new deficiencies were identified.

Conclusion In a real-life cohort of patients with NFPAs, recovery of hypopituitarism after surgery is more frequent than the occurrence of new deficiencies. Hence, hypopituitarism could be considered a relative indication for surgery in patients with NFPAs.

Keywords NFPAs · Transsphenoidal surgery · Pituitary function · Hypopituitarism

✉ Maria Mavromati
maria.mavromati@hcuge.ch

[1] Service of Endocrinology, Diabetes, Nutrition and Therapeutic Patient Education, WHO Collaborating Center, Geneva University Hospital, Geneva University, Geneva, Switzerland

[2] Division of Nephrology, McGill University Health Center, McGill University, Montreal, QC, Canada

[3] Service of Neurosurgery, Geneva University Hospital, Geneva University, Geneva, Switzerland

[4] Service of Neurodiagnostic, Division of Neuroradiology, Geneva University Hospital, Geneva University, Geneva, Switzerland

[5] Service of Clinical Pathology, Geneva University Hospital, Geneva, Switzerland

[6] Geneva University, Geneva, Switzerland

Introduction

Pituitary adenomas are the most common tumors of the sella turcica and are of benign nature. Their incidence in autopsy series reaches 10% of the population. Their prevalence ranges in-between 78 and 94 cases per 100,000 habitants in recent studies, showing that these tumors are not as rare as they were believed to be [1, 2]. Non-functioning pituitary adenomas (NFPAs), defined by the absence of clinical and biological evidence of hormonal secretion, represent 25–40% of all pituitary adenomas [3–5] They are the second most common subtype of pituitary adenomas with prolactinomas being the most prevalent (40–55%). The other subtypes are less frequent, consisting in GH-secreting adenomas in 10% of the cases, ACTH-

secreting adenomas in 1–5% of the cases and TSH-secreting adenomas in less than 1% of the cases [6].

Immunohistochemistry of NFPAs usually shows gonadotropin expression (68%), or no hormonal expression at all (null cell adenomas, 27%), while GH, ACTH, TSH or even prolactin expression is quite rare (silent adenomas, 5%) [7, 8]. With routine use of immunohistochemistry for transcription factors, as recommended by the 2022 World Health Organization (WHO) classification for pituitary tumors, pituitary adenomas are distinguished in PIT1-lineage, TPIT-lineage and SF1-lineage pituitary neuroendocrine tumors (PitNETs), with gonadotroph tumors included in the latter category and null cell tumors having no distinct cell lineage, thus, being a diagnosis of exclusion [9].

Based on their size, pituitary adenomas are classified as microadenomas (≤10 mm) or macroadenomas (>10 mm). Unlike non-functioning (NF) microadenomas, which will rarely grow during follow-up, and among which only 5% will exceed 10 mm in diameter, NF macroadenomas seem to have a higher growth potential. Indeed, 25–50% of these tumors will progress during a median follow-up of 2–7 years [10].

In the absence of surgical treatment, the risk of developing new hormone deficiencies in patients with pituitary macroadenomas is estimated to be 12% per year [11]. For NF macroadenomas, surgery is mainly indicated in case of visual impairment. Guidelines from the Endocrine Society suggest that surgery may also be considered in case of hypopituitarism despite the absence of visual impairment, but the volume of evidence is quite limited [6]. The impact of surgery on anterior pituitary function is not yet very well established and risks of new postoperative pituitary deficiencies and recovery after surgery vary among studies [10]. Serious complications related to transsphenoidal surgery are rare (mortality ≤1% and other non-lethal serious complications ≤5%), but cannot be neglected [12]. It is also necessary to take into account the risk of apoplexy of unoperated pituitary adenomas, which is however low and estimated to be of 1% per year [11]. Thus, in the absence of visual impairment, the utility of surgery is controversial, and the decision must be individualized by weighing potential benefits of the intervention and risk of complications.

The objectives of our study were to evaluate the impact of transsphenoidal surgery on anterior pituitary function in patients with NF macroadenomas and to assess factors predicting postoperative recovery and new deficiencies.

Materials and methods

Population and data

We retrospectively reviewed files from 310 consecutive pituitary surgeries performed in Geneva University Hospital, a tertiary center, from March 2004 until January 2018 and selected patients having had surgery for a NF pituitary macroadenoma, on the basis of clinical and biochemical data as well as pathological analysis of the tumor. Clinically functioning adenomas, Rathke cleft cysts, apoplexy, microadenomas, other types of sellar tumors (e.g., craniopharyngiomas), transcranial approach surgery, second surgery, patients with previous radiotherapy, as well as silent ACTH-secreting and GH-secreting adenomas were excluded (Fig. 1). Finally, the analysis included NF pituitary macroadenomas at the time of their first surgical treatment. We documented time, age, symptoms and signs upon diagnosis and surgery, duration of follow-up until surgery, visual field defects and visual acuity as well as reasons for surgery. The study was approved by the Swiss Ethics Committee in compliance with the Declaration of Helsinki; a waiver of informed consent was granted, as the study was determined to involve no risk to the subjects included by using existing medical file information.

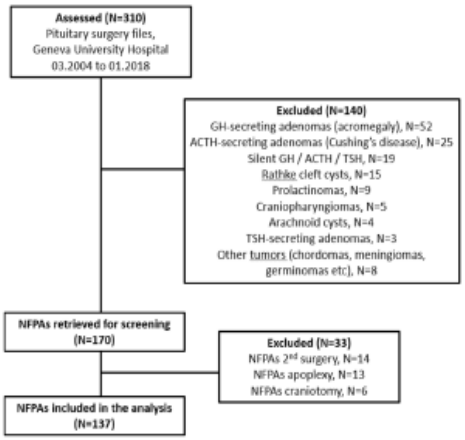

Fig. 1 Flow-chart of pituitary surgeries screened and selected for inclusion in the analysis

Assessment of tumor size

Tumor size was evaluated on gadolinium-enhanced magnetic resonance imaging (MRI), performed 0–3 months before surgery, as well as 3–6 months postoperatively. MRI files were all reviewed in 2-dimension T1 sequences after gadolinium injection and adenomas were measured in 3 diameters (cranio-caudal, antero-posterior and transverse). Patients were classified, before and after surgery, according to the maximal diameter (Dmax) and according to the average diameter (Dav), calculated as the average of 3 diameters, in 4 size groups (S1: 10–19 mm, S2: 20–29 mm, S3: 30–39 mm and S4: ≥40 mm). Finally, we classified

patients according to percentage of decrease of the average diameter with surgery, in groups 0 to 3 (0: ≥80% decrease, 1: 50–79, 2: 20–49 and 3: <20% decrease).

Evaluation of hormonal results

We retrieved and collected the results of hormonal workup performed up to 6 months prior to surgery, as well as 3–6 months postoperatively. Tests had been performed either in Geneva University Hospital, or in private laboratories, and had been organized by the treating physician. Diagnosis of hypopituitarism in our study was based on biologic tests according to normative values suggested by each laboratory, as well as symptoms related to hormonal deficiency, and patients were classified as having a normal or an impaired function, per axis. Percentages of hormonal deficiencies were calculated for patients with available data and not for the whole cohort.

Evaluation of GH deficiency was based on IGF-1 levels, and classification of patients was performed according to age-related normative values suggested by each laboratory.

Assessment of central adrenal insufficiency was based on basal plasma cortisol levels as well as a standard-dose (250 µg) ACTH stimulation test. Patients were classified in five groups as follows: "normal" (if basal cortisol levels were >400 nmol/l or if the level rose >500 nmol/l after ACTH stimulation test), "probably normal" (if dynamic test had not been performed but basal cortisol levels were >350 nmol/l), "central adrenal insufficiency" (if 8 a.m. plasma cortisol levels were <70 nmol/l or <500 nmol/l after dynamic testing) and "possible insufficiency" (if dynamic test had not been performed, 8 a.m. plasma cortisol levels were 70–350 nmol/l, and the patient was receiving glucocorticoid replacement because of symptoms related to possible central adrenal insufficiency). Confounding factors, such as ongoing oral estrogen use, were documented and those patients were thus not classified into the above-mentioned categories. Finally, several patients had an incomplete assessment and could not be classified in the categories described above. On the basis of this classification, two types of analysis were carried out for the ACTH axis evaluation, a strict analysis, comparing only "normal" patients to those with well-documented "central adrenal insufficiency", in line with cut-offs suggested in the 2016 Endocrine Society guidelines for hypopituitarism, as well as a simplified analysis, where "normal" and "probably normal" patients were considered together and compared to patients with "central adrenal insufficiency" and "possible insufficiency" who were also considered together [13].

Assessment of central hypogonadism was based on gender and age. In women of childbearing age, diagnosis of central hypogonadism was based on the presence of oligomenorrhea or amenorrhea, low serum estradiol levels and normal or low FSH/LH levels. In women older than 60 years of age, FSH/LH values at pre-menopausal levels, according to normative values suggested by each laboratory, were sufficient for the diagnosis of central hypogonadism. In men, central hypogonadism was defined by the presence of low total testosterone levels and normal or low LH, together with symptoms or signs of testosterone deficiency. Confounding factors to the evaluation of the gonadal axis were also documented and patients taking hormonal substitution or contraception were not classified.

Diagnosis of central hypothyroidism was based on a low free-T4 together with low or normal TSH. Postoperatively, evaluation was only taken into account if tests had been performed after discontinuation of thyroid hormone replacement if started after surgery. Patients on levothyroxine for primary hypothyroidism were considered as non-assessable.

Patients were classified on the basis of prolactin levels (elevated, normal or low) before and after surgery. To eliminate the bias of the impact of hyperprolactinemia on the gonadal axis, a second analysis of the latter was performed, after excluding patients with hyperprolactinemia.

In order to evaluate clinically relevant global pituitary function improvement or worsening, we performed an additional analysis after excluding cases of GH deficiency (in the absence of routinely performed dynamic testing) as well as cases of hypogonadism in postmenopausal women. Pituitary function was considered improved if at least one axis had improved with a total number of hormonal deficiencies lower than preoperatively. Pituitary function was considered worse if there was at least one new deficiency and a total number of hormonal deficiencies higher than preoperatively.

Surgery

Surgery was performed using a microscopic, endoscope-assisted, trans-septal transsphenoidal approach, by the same experienced surgeon. If patients had postoperative basal cortisol levels below 500 nmol/l on day 4, they were then maintained on 5 mg of prednisone per day until a proper biologic evaluation could take place.

Statistical analysis

Baseline characteristics were reported as mean ± standard deviation, median (interquartile range), or number (percentage), as appropriate. To identify potential factors associated with new hormone deficiencies or the recovery from a previous hormonal deficiency, logistic regression analysis was performed. Five predictors (age per 5 years, gender, tumor size per 5 mm, ≥50% reduction in tumor size after surgery and presence of hyperprolactinemia upon diagnosis) were examined at a univariate level. For the

Table 1 Patients' characteristics

Patients' characteristics	All patients ($N = 137$)
Female/Male, *N* (%)	56 (40.9%)/81 (59.1%)
Age at diagnosis, median in years (range)	59 (28–85)
Age at surgery, median in years (range)	60 (28–86)
Time of follow-up till surgery, median in months (range)	12.7 (3–360)
Reasons leading to diagnosis, *N* (%)	
Incidentaloma	55 (40.1%)
Visual impairment	45 (32.8%)
Symptoms of hormonal dysfunction	23 (16.8%)
Headaches	14 (10.2%)
Visual impairment upon diagnosis, *N* (%)	
Present	80 (58.4%)
Absent	57 (41.6%)
Indication for surgery, *N* (%)	
Visual impairment or optic nerve compression	123 (89.8%)
Tumor growth during follow-up	9 (6.6%)
Other reasons (patient's preference, hypopituitarism, etc.)	5 (3.6)
Tumor size before surgery by maximal diameter (Dmax), $N = 137$	
S1: 10 ≤ Dmax < 20 mm	35 (25.5%)
S2: 20 ≤ Dmax < 30 mm	66 (48.2%)
S3: 30 ≤ Dmax < 40 mm	26 (19%)
S4: Dmax ≥ 40 mm	10 (7.3%)
Tumor size before surgery by average diameter (Dav), $N = 137$	
S1: Dav < 20 mm	55 (40.1%)
S2: 20 ≤ Dav < 30 mm	67 (48.9%)
S3: 30 ≤ Dav < 40 mm	13 (9.5%)
S4: Dav ≥ 40 mm	2 (1.5%)
Tumor size after surgery by average diameter (Dav), $N = 137$	
S0: no residual tumor	50 (36.5%)
S1: Dav < 10 mm	33 (24.1%)
S2: 10 ≤ Dav < 20 mm	30 (21.9%)
S3: Dav ≥ 20 mm	13 (9.5%)
No data	11 (8%)
Average diameter decrease with surgery (Dav), $N = 137$	
S0: >80%	52 (38%)
S1: 50–79%	41 (29.9%)
S2: 20–49%	20 (14.6%)
S3: <20%	12 (8.8%)
No data	12 (8.8%)

outcome of recovery from a previous hormonal deficiency, predictors with a univariate p value < 0.20 were included in the multivariate model. Statistical analyses were performed in SPSS version X and in Stata version 17.0 SE. Any p value < 0.05 was considered significant.

Results

Patients' characteristics

Patients' characteristics are summarized in Table 1. Altogether, 310 pituitary surgery files performed between March 2004 and January 2018 were reviewed, of which 137 cases of non-functioning pituitary macroadenomas were selected. The cohort included 56 women (40.9%) and 81 men (59.1%). Median age at diagnosis was 59 years (range: 28–85) and median age at the time of surgery 60 years (range: 28–86). There were no cases of diabetes insipidus upon diagnosis nor after surgery (3 and 6 months postoperatively). Median time of follow up until surgery was 12.7 months (range: 3–360), while 75% of patients had surgery performed within 8 months from diagnosis and only 14.7% had surgery later than 24 months from diagnosis. On ophthalmologic evaluation following diagnosis (visual field and visual acuity examination), 80 patients (58.4%) had visual impairment. Indication for surgery was visual optic nerve compression on MRI with or without visual impairment in most cases (89.8%, $N = 123$), while 6.6% ($N = 9$) had surgery due to tumor growth during follow-up, and only 3.6% ($N = 5$) for other reasons (hypopituitarism, patient's preference).

Tumor size

Classification of patients according to tumor size before and after surgery is also summarized in Table 1. At the time of surgery, median maximal tumor diameter was 24.8 mm (range: 10–50). In 73.7% of cases ($N = 101$), Dmax was <30 mm (89% with Dav <30 mm). On MRI performed 3–6 months after surgery, 36.5% of patients ($N = 50$) had no residual tumor, 24.1% ($N = 33$) had a residual tumor of <10 mm, while 9.5% ($N = 13$) had a residual tumor exceeding 20 mm of average diameter. Overall, postoperatively, 38% of patients ($N = 52$) had Dav decrease of 80% or more, 29.9% ($N = 41$) had a 50–79% decrease and 23.4% ($N = 32$) had <50% decrease.

Evaluation of the anterior pituitary hormonal axis before and after surgery

Before surgery, central hypogonadism was detected in 62.4% of patients (31.5% of which were postmenopausal women), central hypothyroidism in 41%, central adrenal insufficiency in 30.8% (21.3% with the strict analysis), and low IGF-1 in 29.9% of patients (Table 2 and Fig. 2). High prolactin levels were detected in 50.8% of patients. After surgery, central hypogonadism was detected in 40.8% of patients (30% of which were postmenopausal women), central hypothyroidism in 29.3%, central adrenal

Table 2 Classification of patients according to hormonal workups by axis, before and after surgery

	Prior to surgery	After surgery
LH/FSH		
Number of patients analyzed	117[a]	98[b]
Normal, *n*(%)	44 (37.6%)	58 (59.2%)
Insufficiency, *n*(%)	73 (62.4%)	40 (40.8%)
Insufficiency in postmenopausal women, *n*(%)	23 (19.7%)	12 (12.2%)
TSH		
Number of patients analyzed	122[c]	116[d]
Normal, *n*(%)	72 (59%)	82 (70.7%)
Insufficiency, *n*(%)	50 (41%)	34 (29.3%)
ACTH		
Number of patients analyzed	131[e]	124[f]
Normal, *n*(%)	59 (45%)	75 (60.5%)
Probably normal, *n*(%)	13 (9.9%)	7 (5.6%)
Insufficiency, *n*(%)	16 (12.2%)	22 (17.7%)
Possible insufficiency, *n*(%)	16 (12.2%)	9 (7.3%)
Inconclusive data, *n*(%)	5 (3.8%)	4 (3.2%)
Confounding factors, *n*(%)	22 (16.8%)	7 (5.6%)
ACTH (simplified analysis)		
Number of patients analyzed	104[e]	113[f]
Normal + Probably normal, *n*(%)	72 (69.2%)	82 (72.6%)
Insufficiency + Possible insufficiency, *n*(%)	32 (30.8%)	31 (27.4%)
ACTH (strict analysis)		
Number of patients analyzed	75[e]	97[f]
Normal, *n*(%)	59 (78.7%)	75 (77.3%)
Insufficiency, *n*(%)	16 (21.3%)	22 (22.7%)
GH		
Number of patients analyzed	117[g]	88[h]
Normal, *n*(%)	82 (70.1%)	73 (83%)
Insufficiency, *n*(%)	35 (29.9%)	15 (17%)
Prolactin		
Number of patients analyzed	130[i]	89[j]
High, *n*(%)	66 (50.8%)	12 (13.5%)
Normal, *n*(%)	63 (48.5%)	75 (84.3%)
Low, *n*(%)	1 (0.8%)	2 (2.2%)

[a]No data in 16 cases, confounding factors in 4 cases
[b]No data in 37 cases, confounding factors in 2 cases
[c]No data in 7 cases, confounding factors in 8 cases
[d]No data in 17 cases, confounding factors in 4 cases
[e]No data in 6 cases
[f]No data in 13 cases
[g]No data in 20 cases
[h]No data in 49 cases
[i]No data in 7 cases
[j]No data in 48 cases

insufficiency in 27.4% (22.7% with the strict analysis), and low IGF-1 in 17% of patients (Table 2 and Fig. 2). High prolactin level was detected in 13.5% of patients.

Among 122 patients with data available before and after surgery, 42 (34.4%) showed recovery of at least one deficiency and 13 (10.6%) experienced the occurrence of one or more new hormonal deficiencies (Fig. 3). Gonadotropin secretory reserve recovered in 35.7% (20/56) of patients (5 of which were postmenopausal women, 25%) while new-onset central hypogonadism occurred in 8.3% (3/36) of patients (none of which were postmenopausal women). Those numbers were 27.3% and 5.3%, respectively for the 41 patients without hyperprolactinemia upon diagnosis. Thyrotropin secretory reserve recovered in 30.4% (14/46) of cases, while new-onset central hypothyroidism occurred in 1.6% (1/62) of patients. With the simplified analysis of the ACTH axis, ACTH axis recovered in 15.4% (4/26) of patients while new-onset central adrenal insufficiency occurred, in 9.2% of patients. With the strict analysis of ACTH-secretion evaluation, ACTH secretory reserve recovered in 40% of them (4/10), while new-onset central adrenal insufficiency occurred in 9.6% of patients (5/52). Recovery of the GH axis occurred in 45.5% (10/22) of patients while new-onset GH deficiency after surgery occurred in 5.1% (3/59). Finally, prolactin levels normalized in 71.4% of cases (30/42), while no cases of new hyperprolactinemia were detected after surgery.

In an additional analysis evaluating global pituitary function change with surgery and taking into account only clinically relevant deficiencies requiring treatment, we found 30 patients with global improvement of pituitary function (24.6%) and 7 with global worsening (5.7%), among 122 with data available before and after surgery in at least one axis.

Predictors of new deficiencies and recovery after surgery

In the univariate analysis, there was a trend toward increased chances of recovery of at least one deficient hormonal axis in patients with hyperprolactinemia, male gender and older patients which did not reach statistical significance (Table 3). Tumor size was not associated with recovery. In multivariate analysis, hyperprolactinemia and male gender were confirmed to predict postoperative recovery of at least one deficient hormonal axis (Table 3).

On the contrary, in univariate analysis, gender, age, tumor size and hyperprolactinemia, were not found to independently predict the appearance of new hormonal deficiencies after surgery, and thus, a multivariate analysis was not performed (Table 3).

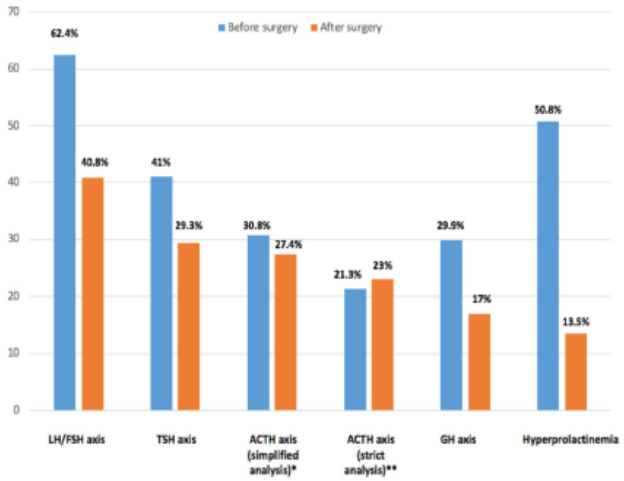

Fig. 2 Rates of hormonal insufficiencies per axis and hyperprolactinemia, before and after surgery; patients with data before or after surgery. *Simplified analysis for the ACTH axis: patients with morning cortisol levels >350 nmol/l or >500 nmol/l after ACTH stimulation test were considered as normal, patients with morning cortisol levels <70 nmol/l or <500 nmol/l after dynamic testing as well as patients with morning cortisol levels 70–350 nmol/l but who were receiving glucocorticoid replacement because of symptoms related to possible adrenal insufficiency, were considered as having central adrenal insufficiency. **Strict analysis for the ACTH axis: only patients with morning cortisol levels >400 nmol/l or >500 nmol/l after ACTH stimulation test were considered as normal, only patients with morning cortisol levels <70 nmol/l or <500 nmol/l after dynamic testing were considered as having central adrenal insufficiency

Discussion

In the absence of threat for the optic pathways, surgery for non-functioning pituitary macroadenoma is not systematically recommended. Nevertheless, if a potential improvement of pituitary function with surgery is aimed at, it would be useful to have prognostic factors of better outcome in order to be able to select patients who could benefit from the surgical excision in the absence of visual impairment.

The impact of hypopituitarism, whether preoperative or postoperative, on patients' long-term outcomes, must be taken into account in order to assess potential benefits of surgery. Current data suggest higher morbidity and mortality in patients with NF pituitary macroadenoma, and hypopituitarism is found to be a risk factor, regardless of it being the result of surgery or occurring during the simple surveillance of a growing adenoma [14]. A Danish registry-based study on 192 patients who had surgery for a NFPA did not find any increase in mortality (standardized mortality ratio—SMR: 1.21, 95% CI: 0.93–1.59), regardless of the presence of hypopituitarism [15]. On the other hand, the largest study on the topic, based on a Swedish registry including 2795 patients with NFPA (among whom 52% had surgery), found, after follow-up of a median of 7 years, a small but significant increase in mortality (SMR: 1.10; 95% CI: 1.00–1.20). Mortality was significantly higher in patients who were younger than 40 years old (SMR: 2.68, 95% CI: 1.23–5.09), and in women (SMR: 1.29, 95% CI: 1.11–1.48), and was attributed to cerebrovascular and infectious causes [16]. In a British registry-based study including 546 patients who had surgery for NFPA, increased mortality was found, after follow-up of a median of 8 years (SMR: 3.5; 95% CI: 2.8–4.4). In this cohort, the only independent predictor of mortality was age at diagnosis (hazard ratio (HR): 1.1 if >50 years, 95% CI: 1.07–1.13, $p < 0.001$) [17]. Furthermore, hypopituitarism requiring treatment seems to contribute to morbidity of patients with NFPA. Treatment for central adrenal insufficiency for example requires adjustment of glucocorticoids and overtreatment has been associated to increased mortality [18]. In addition, patients with growth hormone deficiency after surgical management for a NFPA and who do not receive treatment with GH, seem to have a higher risk of developing type 2 diabetes (odds ratio (OR): 1.65, 95% CI: 1.06–2.46, $p = 0.018$) [19]. A British population study on 519 patients (90.6% operated, 9.4% observed) with a follow-up of 7 years, showed that central adrenal insufficiency and central hypogonadism were associated with increased mortality (relative risk (RR): 2.26; 95% CI: 1.15–4.47, and 2.56; 95% CI: 1.10–5.96 respectively). In this study, there was a trend toward higher mortality rates

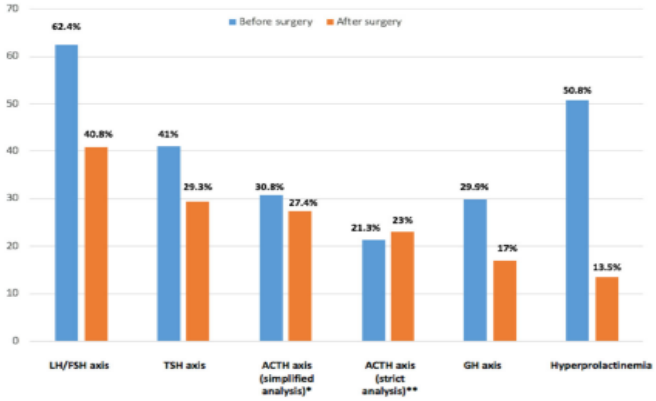

Fig. 2 Rates of hormonal insufficiencies per axis and hyperprolactinemia, before and after surgery; patients with data before or after surgery. *Simplified analysis for the ACTH axis: patients with morning cortisol levels >350 nmol/l or >500 nmol/l after ACTH stimulation test were considered as normal, patients with morning cortisol levels <70 nmol/l or <500 nmol/l after dynamic testing as well as patients with morning cortisol levels 70–350 nmol/l but who were receiving glucocorticoid replacement because of symptoms related to possible adrenal insufficiency, were considered as having central adrenal insufficiency. **Strict analysis for the ACTH axis: only patients with morning cortisol levels >400 nmol/l or >500 nmol/l after ACTH stimulation test were considered as normal, only patients with morning cortisol levels <70 nmol/l or <500 nmol/l after dynamic testing were considered as having central adrenal insufficiency

Discussion

In the absence of threat for the optic pathways, surgery for non-functioning pituitary macroadenoma is not systematically recommended. Nevertheless, if a potential improvement of pituitary function with surgery is aimed at, it would be useful to have prognostic factors of better outcome in order to be able to select patients who could benefit from the surgical excision in the absence of visual impairment.

The impact of hypopituitarism, whether preoperative or postoperative, on patients' long-term outcomes, must be taken into account in order to assess potential benefits of surgery. Current data suggest higher morbidity and mortality in patients with NF pituitary macroadenoma, and hypopituitarism is found to be a risk factor, regardless of it being the result of surgery or occurring during the simple surveillance of a growing adenoma [14]. A Danish registry-based study on 192 patients who had surgery for a NFPA did not find any increase in mortality (standardized mortality ratio—SMR: 1.21, 95% CI: 0.93–1.59), regardless of the presence of hypopituitarism [15]. On the other hand, the largest study on the topic, based on a Swedish registry including 2795 patients with NFPA (among whom 52% had surgery), found, after follow-up of a median of 7 years, a small but significant increase in mortality (SMR: 1.10; 95% CI: 1.00–1.20). Mortality was significantly higher in patients who were younger than 40 years old (SMR: 2.68, 95% CI: 1.23–5.09), and in women (SMR: 1.29, 95% CI: 1.11–1.48), and was attributed to cerebrovascular and infectious causes [16]. In a British registry-based study including 546 patients who had surgery for NFPA, increased mortality was found, after follow-up of a median of 8 years (SMR: 3.5; 95% CI: 2.8–4.4). In this cohort, the only independent predictor of mortality was age at diagnosis (hazard ratio (HR): 1.1 if >50 years, 95% CI: 1.07–1.13, $p < 0.001$) [17]. Furthermore, hypopituitarism requiring treatment seems to contribute to morbidity of patients with NFPA. Treatment for central adrenal insufficiency for example requires adjustment of glucocorticoids and overtreatment has been associated to increased mortality [18]. In addition, patients with growth hormone deficiency after surgical management for a NFPA and who do not receive treatment with GH, seem to have a higher risk of developing type 2 diabetes (odds ratio (OR): 1.65, 95% CI: 1.06–2.46, $p = 0.018$) [19]. A British population study on 519 patients (90.6% operated, 9.4% observed) with a follow-up of 7 years, showed that central adrenal insufficiency and central hypogonadism were associated with increased mortality (relative risk (RR): 2.26; 95% CI: 1.15–4.47, and 2.56; 95% CI: 1.10–5.96 respectively). In this study, there was a trend toward higher mortality rates

Table 3 Factors predicting pituitary function improvement and worsening

	Factors predicting recovery of at least 1 hormonal axis ($N = 42/122$)				Factors predicting appearance of at least 1 new hormonal deficiency ($N = 13/122$)	
	Univariate analysis		Multivariate analysis		Univariate analysis	
	OR (95% CI)	*p* value	OR (95% CI)	*p* value	OR (95% CI)	*p* value
Gender (female)	0.51 (0.23–1.15)	0.104	0.27 (0.10–0.71)	0.009	0.997 (0.31–3.25)	0.996
Age (per 5 years)	1.12 (0.98–1.30)	0.106	1.14 (0.98–1.34)	0.095	1.16 (0.92–1.45)	0.214
Tumor size prior to surgery—maximal diameter (per 5 mm)	0.95 (0.75–1.19)	0.637			1.06 (0.75–1.49)	0.749
Average diameter decrease with surgery (at least 50% decrease)	1.07 (0.45–2.56)	0.882			0.79 (0.22–2.77)	0.712
Hyperprolactinemia	2.09 (0.96–4.57)	0.064	3.45 (1.38–8.64)	0.008	1.57 (0.48–5.12)	0.455

multivariate regression analysis model including gender and age. It is difficult to explain the mechanism underlying this association. One reason, suggested by the authors of the French and Belgian cohort [24], could be the fact that pituitary stalk compression, rather than destruction of pituitary cells, is the possible cause of hypopituitarism in patients with hyperprolactinemia, and can thus be reversible with surgery.

As for the risk of new postoperative deficiencies, only a few studies showed higher risk in patients with larger tumors [28, 30]. This finding was not confirmed by our study; indeed, we did not identify any predictive factor of pituitary function worsening with surgery, however, this finding must be interpreted with caution since sample size for this analysis was very small (only 13 patients, among 122 with data available before and after surgery, with at least one new hormonal deficiency).

The main limitations of our study are related to its retrospective design. Missing data were present, especially in hormonal axis considered "non clinically relevant", as the gonadal axis in women after menopause. Nevertheless, it represents real-life practice and underlines the importance of a rigorous hormonal evaluation. Indeed, we discovered that only 54.8% of patients had a complete workup of ACTH secretory reserve upon diagnosis and 70.8% after surgery; those patients had lower rates of central adrenal insufficiency compared to patients with a less detailed evaluation, thus avoiding the continuation of unnecessary glucocorticoid replacement therapy. Evaluation of GH deficiency by means of serum IGF-1 lacks sensitivity since about 20% of adults with GH deficiency have normal IGF-1 levels, still, dynamic testing is not routinely performed in real-life practice in asymptomatic patients. The assessment of tumor size remains imperfect, especially in the presence of bilateral residual tumors, with the risk of underestimating the result of surgery. Finally, our results come from a tertiary center and should be generalized with caution; less experienced centers may face increased rates of complications and less favorable outcomes.

Conclusion

Following transsphenoidal surgery for non-functioning pituitary macroadenoma the rate of at least one new postoperative hormonal deficiency was lower than the recovery rate of at least one hormonal axis. The ACTH secretory reserve was the most fragile while TSH secretory reserve seemed to be the most resistant. Men were more likely to recover from a preexisting central hormonal deficiency after surgery, yet, the presence of hyperprolactinemia was found to be the strongest predictor of pituitary function recovery. Based on the encouraging recovery rate along with the

relatively low risk of new deficiencies, hypopituitarism due to a NFPA could be considered as a valid relative indication for surgery.

Funding Open access funding provided by University of Geneva.

Compliance with ethical standards

Conflict of interest The authors declare no competing interests.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

References

1. A.F. Daly et al. High prevalence of pituitary adenomas: a cross-sectional study in the province of Liege, Belgium. J. Clin. Endocrinol. Metab. **91**(12), 4769–4775 (2006)
2. A. Fernandez, N. Karavitaki, J.A. Wass, Prevalence of pituitary adenomas: a community-based, cross-sectional study in Banbury (Oxfordshire, UK). Clin. Endocrinol. **72**(3), 377–382 (2010)
3. P. Chanson et al. Management of clinically non-functioning pituitary adenoma. Ann. Endocrinol. **76**(3), 239–247 (2015)
4. H. Buurman, W. Saeger, Subclinical adenomas in postmortem pituitaries: classification and correlations to clinical data. Eur. J. Endocrinol. **154**(5), 753–758 (2006)
5. L.N. Vieira et al. A review on the diagnosis and treatment of patients with clinically nonfunctioning pituitary adenoma by the Neuroendocrinology Department of the Brazilian Society of Endocrinology and Metabolism. Arch. Endocrinol. Metab. **60**(4), 374–390 (2016)
6. P.U. Freda et al. Pituitary incidentaloma: an endocrine society clinical practice guideline. J. Clin. Endocrinol. Metab. **96**(4), 894–904 (2011)
7. W. Saeger et al. Pathohistological classification of pituitary tumors: 10 years of experience with the German Pituitary Tumor Registry. Eur. J. Endocrinol. **156**(2), 203–216 (2007)
8. A.F. Daly, A. Beckers, The epidemiology of pituitary adenomas. Endocrinol. Metab. Clin. North Am. **49**(3), 347–355 (2020)
9. S.L. Asa et al. Overview of the 2022 WHO Classification of Pituitary Tumors. Endocr. Pathol. **33**(1), 6–26 (2022)
10. O.M. Dekkers, A.M. Pereira, J.A. Romijn, Treatment and follow-up of clinically nonfunctioning pituitary macroadenomas. J. Clin. Endocrinol. Metab. **93**(10), 3717–3726 (2008)
11. F. Castinetti et al. Non-functioning pituitary adenoma: when and how to operate? What pathologic criteria for typing?. Ann. Endocrinol. **76**(3), 220–227 (2015)
12. M.H. Murad et al. Outcomes of surgical treatment for non-functioning pituitary adenomas: a systematic review and meta-analysis. Clin. Endocrinol. **73**(6), 777–791 (2010)
13. M. Fleseriu et al. Hormonal replacement in hypopituitarism in adults: an endocrine society clinical practice guideline. J. Clin. Endocrinol. Metab. **101**(11), 3888–3921 (2016)
14. M. Tampourlou et al. Mortality in patients with non-functioning pituitary adenoma. Pituitary **21**(2), 203–207 (2018)
15. E.H. Nielsen et al. Nonfunctioning pituitary adenoma: incidence, causes of death and quality of life in relation to pituitary function. Pituitary **10**(1), 67–73 (2007)
16. D.S. Olsson et al. Excess mortality in women and young adults with nonfunctioning pituitary adenoma: a Swedish Nationwide Study. J. Clin. Endocrinol. Metab. **100**(7), 2651–2658 (2015)
17. G. Ntali et al. Mortality in patients with non-functioning pituitary adenoma is increased: systematic analysis of 546 cases with long follow-up. Eur. J. Endocrinol. **174**(2), 137–145 (2016)
18. T. Zueger et al. Glucocorticoid replacement and mortality in patients with nonfunctioning pituitary adenoma. J. Clin. Endocrinol. Metab. **97**(10), E1938–E1942 (2012)
19. C. Hammarstrand et al. Comorbidities in patients with non-functioning pituitary adenoma: influence of long-term growth hormone replacement. Eur. J. Endocrinol. **179**(4), 229–237 (2018)
20. M.W. O'Reilly et al. ACTH and gonadotropin deficiencies predict mortality in patients treated for nonfunctioning pituitary adenoma: long-term follow-up of 519 patients in two large European centres. Clin. Endocrinol. **85**(5), 748–756 (2016)
21. J.A. Gondim et al. Endoscopic endonasal approach for pituitary adenoma: surgical complications in 301 patients. Pituitary **14**(2), 174–183 (2011)
22. J.H. Kim et al. Endoscopic transsphenoidal surgery outcomes in 331 nonfunctioning pituitary adenoma cases after a single surgeon learning curve. World Neurosurg. **109**, e409–e416 (2018)
23. M. Messerer et al. Evidence of improved surgical outcome following endoscopy for nonfunctioning pituitary adenoma removal. Neurosurg. Focus **30**(4), E11 (2011)
24. O. Alexopoulou et al. Outcome of pituitary hormone deficits after surgical treatment of nonfunctioning pituitary macroadenomas. Endocrine **73**(1), 166–176 (2021)
25. M. Messerer et al. Non-functioning pituitary macro-incidentalomas benefit from early surgery before becoming symptomatic. Clin. Neurol. Neurosurg. **115**(12), 2514–2520 (2013)
26. S.M. Webb et al. Recovery of hypopituitarism after neurosurgical treatment of pituitary adenomas. J. Clin. Endocrinol. Metab. **84**(10), 3696–3700 (1999)
27. P. Nomikos et al. Impact of primary surgery on pituitary function in patients with non-functioning pituitary adenomas—a study on 721 patients. Acta Neurochir. **146**(1), 27–35 (2004)
28. N. Fatemi et al. Pituitary hormonal loss and recovery after transsphenoidal adenoma removal. Neurosurgery **63**(4), 709–718 (2008).
29. A. Jahangiri et al. Improved versus worsened endocrine function after transsphenoidal surgery for nonfunctional pituitary adenomas: rate, time course, and radiological analysis. J. Neurosurg. **124**(3), 589–595 (2016)
30. J.Y. Hwang et al. Axis-specific analysis and predictors of endocrine recovery and deficits for non-functioning pituitary adenomas undergoing endoscopic transsphenoidal surgery. Pituitary **23**(4), 389–399 (2020)

ii. Reference values for IGF-1 serum concentrations: comparison of six immunoassays (VARIETE Study)

Reliable results of IGF-1 measurement are essential for diagnosis and follow-up of patients with growth hormone (GH) disorders, either acromegaly or GH deficiency. However, commercially available immunoassays give considerably different results for the same sample even if consensus recommendations for standardization and evaluation of IGF-1 assays are followed. Inter-assay variability is due to calibration against different IGF-1 standards and different methods for IGF-bind proteins (IGFBPs) removal. In addition, normal values for IGF-1 are very difficult to establish since they depend on age, sex, BMI, estrogen or testosterone replacement treatments, renal and liver function as well as the presence of metabolic disease mostly diabetes mellitus. Thus, different normal values are suggested by different fabricants for each assay, but also by different laboratories using the same assay.

The VARIETE Study (VAleurs de Référence de l'IGF-I Et Transformation En z score) was a cross-sectional multicenter study with the objective to establish reference values for six widely used immunoassays (iSYS, LIAISON XL, IMMULITE, IGFI RIACT, Mediagnost ELISA and Mediagnost RIA) in a healthy French population. The study included 911 adults (18-90 years, 470 males), with approximately 100 subjects per decade, from 10 French centers. Inclusion was done after a thorough medical history, clinical examination and biological check-up, to exclude individuals with medications or health problems influencing IGF-1.

Since IGF-1 raw values distributions are non-Gaussian, they were normalized with Box-Cox power transformation, then, age- and sex-specific curves were established, by means of the LMS method (L: skewness, M: median, S: coefficient of variation) to allow percentile and standard deviation scores (SDS) calculation. Normal reference values (age- and sex-specific) ranged from percentile 2.5 to 97.5. A formula for SDS calculation for each assay, using the name of the assay, raw IGF-1 value, sex and age, was provided, and is available online, with the aim to facilitate patient classification.

For all six assays, IGF-1 levels were found to be significantly higher in women until the age of 59, then were slightly higher in men but not significantly for all assays. Lower limits of defined normative values were similar among all 6 assays but there was considerable variability in upper limits. Pairwise correlation between assays were evaluated and were found to be moderate to good (0.38-0.70). Overall agreement was moderate (Kappa coefficient: 0.55). Correlation with reference values provided by each manufacturer were poor.

The VARIETE Study showed that even when obtained in the same healthy population, reference normative IGF-1 values vary among six widely used immunoassays, which show moderate to good concordance. These findings suggest that IGF-1 monitoring in each patient should ideally be performed with the same assay. Age- and sex-specific reference normative values should be obtained in a large healthy population and calculation of SDS for classification of IGF-1 levels to normal, low or high levels can only be performed after normalization of raw data distribution.

Reference Values for IGF-I Serum Concentrations: Comparison of Six Immunoassays

Philippe Chanson, Armelle Arnoux, Maria Mavromati, Sylvie Brailly-Tabard, Catherine Massart, Jacques Young, Marie-Liesse Piketty, and Jean-Claude Souberbielle for the VARIETE Investigators*

Service d'Endocrinologie et des Maladies de la Reproduction and Centre de Référence des Maladies Endocriniennes Rares de la Croissance (P.C., M.M., J.Y.), Unité de Recherche Clinique (A.A.), and Service de Génétique Moléculaire, Pharmacogénétique et Hormonologie (S.B.T.), Assistance Publique-Hôpitaux de Paris, Hôpitaux Universitaires Paris-Sud, Hôpital de Bicêtre, Le Kremlin-Bicêtre, F94275, France; Inserm 1185 (P.C., S.B.T., J.Y.), Fac Med Paris Sud, Université Paris-Saclay, Le Kremlin-Bicêtre, F-94276, France; and Laboratoire d'Hormonologie (C.M.), Centre Hospitalier Universitaire de Rennes, Centre d'Investigation Clinique Plurithématique, Inserm 1414, Hôpital Pontchaillou, Rennes, F29000, France; Service des Explorations Fonctionnelles (M-L.P., J.-C.S.), Assistance Publique-Hôpitaux de Paris, Hôpital Necker-Enfants Malades, Paris, F75015, France

Context: Measurement of IGF-I is essential for diagnosis and management of patients with disorders affecting the somatotropic axis. However, even when IGF-I kit manufacturers follow recent consensus guidelines, different kits can give very different results for a given sample.

Objectives: We sought to establish normative data for six IGF-I assay kits based on a large random sample of the French general adult population.

Subjects and Methods: In a cross-sectional multicenter cohort study, we measured IGF-I in 911 healthy adults (18–90 years) with six immunoassays (iSYS, LIAISON XL, IMMULITE, IGFI RIACT, Mediagnost ELISA, and Mediagnost RIA). Pairwise concordance between assays was assessed with Bland-Altman plots for both IGF-1 raw data and standard deviation scores (SDS), as well as with the percentage of observed agreement and the weighted Kappa coefficient for categorized IGF-I SDS.

Results: Normative data included the range of values (2.5–97.5 percentiles) given by the six IGF-I assays according to age group and sex. A formula for SDS calculation is provided. Although the lower limits of the reference intervals of the six assays were similar, the upper limits varied markedly. Pairwise concordances were moderate to good (0.38–0.70).

Conclusion: Despite being obtained in the same healthy population, the reference intervals of the six commercial IGF-1 assay kits showed noteworthy differences. Agreement between methods was moderate to good. (*J Clin Endocrinol Metab* 101: 3450–3458, 2016)

Growth hormone exerts its effects on target tissues either directly or via the production of insulin-like growth factor 1 (IGF-I). Accurate measurement of IGF-I in serum is crucial for diagnosis and management of disorders affecting the somatotropic axis, particularly GH excess (acromegaly) and GH deficiency. However, even if manufacturers follow the recommendations of the Consensus Group on the Standardization and Evaluation of GH and IGF-I Assays (1), the different commercial IGF-I assay kits can give very different results for the same sample, with up to a 2.5-fold difference between the lowest and highest values (2). This intermethod variability is generally explained by calibration against different IGF-I reference preparations (3), and differences in the efficiency of

ISSN Print 0021-972X ISSN Online 1945-7197
Printed in USA
Copyright © 2016 by the Endocrine Society
Received January 31, 2016. Accepted May 6, 2016.
First Published Online May 11, 2016

Abbreviations: BMI, body mass index; IGFBP, IGF binding protein; IRR, international reference reagent; LC, liquid chromatography; LMS, parameters L for skewness, M for median, and S for the coefficient of variation; MS, mass spectometry; SDS, standard deviation score.

Downloaded from https://academic.oup.com/jcem/article/101/9/3450/2806775 by guest on 04 May 2025

methods used to remove IGF-binding proteins (IGFBPs) (4). In theory, this should not be a problem in clinical practice because kits that give higher values should have higher normal limits, and patients should thus be consistently classified.

However, it is very difficult to establish reference values for IGF-I. Indeed, serum IGF-I concentrations increase with children's age and pubertal stage, whereas they fall with age in adults (5). Furthermore, the distribution of IGF-I values in an apparently healthy population is non-Gaussian, and this necessitates complex mathematical transformation to obtain reference intervals for each age group. For this reason, it is essential to generate reference values after stratifying a large healthy population into age groups. Another problem is that IGF-I concentrations are influenced by many factors other than GH concentrations, including nutritional status and body mass index (BMI), use of hormone replacement therapy by postmenopausal women, depending on the administration route (6–8), kidney and liver function, and diabetic status (9). Reference IGF-I values may therefore be influenced by the inclusion criteria used to select the reference population sample. This could have important implications for diagnosis and therapeutic decision-making because a given patient could be classified as having a normal IGF-I concentration with one method but an abnormal value with another method. Several studies suggest that the main reason for interlaboratory variability in patient classification is the use of different populations to establish reference values for the different IGF-I assays (2, 10, 11). It is currently difficult to monitor an individual patient with different IGF-I assays, even if the results are all expressed in the same units (ng/ml). It is thus recommended to establish specific reference ranges for each assay and to apply common, well-defined inclusion criteria to the reference population (1). It is also recommended, for the comparison of values obtained with different assays in the same patient, to express each IGF-I result as an SD score (SDS) with reference to the normative data for the assay in question, after appropriate transformation for data non normality. We reasoned that the best way to overcome this variability would be to apply all the commercial kits used in clinical laboratories to a battery of samples from the same well-defined reference population, and to use the same mathematical transformation to calculate reference ranges from the raw data.

The aim of this study was thus to establish normative data for six commercial IGF-I assays in a large random sample of healthy subjects from the French general population representing all adult age groups (about 100 subjects per decade), as recommended by the Consensus Group on the Standardization and Evaluation of GH and IGF-I assays (1). Serum samples from the reference population were tested with six commercial assay kits available in France at the time of this study, after careful exclusion of subjects with medical conditions or medications that might affect their IGF-I concentration. The data were analyzed to obtain the range (2.5–97.5 percentiles) in mass units. The standard deviation scores were used to compare the six assays.

Subjects and Methods

IGF-I assay characteristics

Six immunoassays (iSYS, LIAISON XL, IMMULITE, IGFI RIACT, Mediagnost ELISA, and Mediagnost RIA) were used to measure the IGF-I concentration in each healthy subject. The main characteristics of the assays, and the mathematical models used to determine normative data, where relevant (12–14) as provided by the manufacturer, are shown in Table 1.

Healthy subjects

The subjects were part of a large cohort of French healthy adults (VARIETE). The VARIETE cohort was an open, prospective, national, multicenter, nonrandomized study of healthy volunteers, designed to establish normative data for IGF-I and other hormones in the French general adult population representing all age groups (about 100 subjects per decade from 18–90 years) (ClinicalTrials.gov Identifier: NCT01831648). A total of 972 healthy subjects with BMI values between 19 and 28 kg/m^2 were recruited in 10 centers throughout France between 2010 and 2011. Our objective of including 1000 subjects was not achieved because of difficulties for obtaining an accurate number of subjects in the older age categories (>70 years) fulfilling all the inclusion criteria and without exclusion criteria before the end of our inclusion period. Subjects with medical conditions or medications that might affect IGF-I serum levels were excluded (Supplemental Appendix). Each subject had a clinical examination, personal medical history-taking, and general examination, including careful evaluation of nutritional and gonadal status. Standard laboratory tests (plasma sodium, potassium, calcium, phosphate and creatinine, glycemia, total cholesterol, liver enzymes, TSH, blood cell count, albuminemia, prothrombin time, as well as HIV and hepatitis C virus serologies) were then performed, and 80 ml of blood (50 ml without anticoagulant and 30 ml in EDTA-containing tubes) was sampled and promptly centrifuged (2000 × *g*, 4 C). Serum and plasma were aliquoted, frozen, and stored at −80 C until hormone measurements.

All healthy subjects gave their written informed consent to participate in the study, which was approved by the Paris-Sud Ethics committee before the beginning of the study.

Statistical methods

The distribution of IGF-I values obtained with each assay was skewed, and was thus first normalized by means of sex- and age-specific Box-Cox power transformation. Student's *t* test and Levene's test were then used to assess equality of means and homogeneity of variances between men and women in each age group. As men and women had significantly different IGF-I levels, centile curves were constructed separately for each sex.

Downloaded from https://academic.oup.com/jcem/article/101/9/3450/2806775 by guest on 04 May 2025

Table 1. Characteristics of the Tested IGF-I Assays as Provided by the Manufacturers

Assay Name	Manufacturer	Automated	Tracer	International Standard Against Which the Assay Calibrated	Intra-assay CV	Inter-assay CV	LOQ or LOD (ng/ml)	Highest Measurable Value Without Dilution (ng/ml)	Reference Adult Population Recruited by the Manufacturer
iSYS	IDS	Yes	Acridinium ester	WHO/NIBSC 02/254	2.9% at 22 ng/ml 1.9% at 163 ng/ml 4.2% at 304 ng/ml	5.4% at 22 ng/ml 3.9% at 163 ng/ml 7.2% at 304 ng/ml	8.8 (LOQ)	1200	6500 adults; reference values provided according to the method of Cole and Green (12)
LIAISON XL	DiaSorin	Yes	Isoluminol	WHO/NIBSC 02/254	5.1% at 70 ng/ml 3.5% at 183 ng/ml 3% at 589 ng/ml	9.6% at 80 ng/ml 7.1% at 187 ng/ml 5.6% at 317 ng/ml	3 (LOD) 10 (LOQ)	1500	1606 adults; reference values provided by age according to the method of Royston and Wright (14)
IMMULITE 2000	Siemens	Yes	Alkaline phosphatase	WHO/NIBSC First IRR 87/518	3.9% at 77 ng/ml 6.5% at 169 ng/ml 2.9% at 380 ng/ml 3.0% at 689 ng/ml 2.3% at 1053 ng/ml 2.4% at 1358 ng/ml	7.7% at 77 ng/ml 5.4% at 169 ng/ml 7.4% at 380 ng/ml 8.1% at 689 ng/ml 3.7% at 1053 ng/ml 4.7% at 1358 ng/ml	20 (LOQ)	1600	1499 pediatric and adult samples from an apparently healthy population (no indication is given concerning the respective numbers of adult and children)
IGFI-RIACT	Cisbio	No	^{125}I	WHO/NIBSC First IRR 87/518	3.8% at 49 ng/ml 3.4% at 162 ng/ml 3.2% at 496 ng/ml	3.8% at 39 ng/ml 8.2% at 352 ng/ml 5.9% at 509 ng/ml	1 (LOD)	900	693 adults 29–70 y
Mediagnost	MEDIA	No	Peroxidase enzyme conjugate	WHO/NIBSC 02/254	5.7% at 138 ng/ml	6.1% at 142 ng/ml	1.9 (LOD)	1050	Based on the data reported by Blum and Breier (13)
ELISA	GNOST				5.1% at 141 ng/ml 6.6% at 145 ng/ml	6.8% at 174 ng/ml 2.2% at 494 ng/ml			
Mediagnost	MEDIA	No	^{125}I	WHO/NIBSC 02/254	4.6% at 56 ng/ml	4.9% at 55 ng/ml	2.6 (LOD)	780	Based on the data reported by Blum and Breier (13)
RIA	GNOST				3.4% at 140 ng/ml 2.5% at 180 ng/ml	6.2% at 140 ng/ml 4.5% at 186 ng/ml			The reference values for the different age ranges are the same as those used for the Mediagnost ELISA kit

Abbreviations: CV, coefficient of variation; LOD, limit of detection; LOQ, limit of quantification; NICSC, National Institute for Biological Standards and Control; WHO, World Health Organization.

These six assays are sandwich assays that use monoclonal antibodies directed against epitopes, whose exact nature is not disclosed by the manufacturers. In all cases, IGFBPs are said to be removed by displacement of endogenous IGF-I by an excess of IGF-II (or analog) as initially proposed by Blum and Breier (13). The LOQ is the lowest amount of IGF-I that can be accurately quantified with an allowable error <20%. The LOD is the IGF-I concentration corresponding to the 95th percentile value from a number of determinations of IGF-I concentration in free serum samples.

Age- and sex-specific centile curves were constructed for each assay by using the LMS (parameters L for skewness, M for median, and S for the coefficient of variation) method (12) implemented in the GAMLSS software package version 4.3–1 (15) of R software, version 3.1.2 (R Core Team, 2014; R: A language and environment for statistical computing; R Foundation for Statistical Computing; http://www.R-project.org/.). The LMS method enables smooth curves to be estimated for percentiles after normalization (by Box-Cox power transformation) and standardization of the data. The parameters L, M, and S were also computed for each age and sex class. SDS were calculated as $z = [(\text{IGF-I}/M)^{L} - 1]/(L \times S)$, where IGF-I is the raw value given by the assay (in ng/ml). For each technique, SDS were categorized as low, normal, or high according to their positions relative to both the 2.5th and 97.5th percentiles.

Once the L, M, and S parameters for each category of age and sex had been obtained, the lower and upper reference interval limits were determined for each assay by fixing z at –1.96 and 1.96, respectively, and then mathematically back-transforming the SDS formula.

Pairwise concordance between assays was assessed with scatter plots and Bland-Altman plots for both IGF-I raw values and SDS values, as well as with the percentage of observed agreement (total number of agreements divided by the total number of patients tested with both assays) and the linearly weighted Kappa coefficient for categorized IGF-I SDS (16, 17). An overall kappa coefficient (16) and Friedman's test were computed for global comparison of all assays at the same time. Landis and Koch's table was followed for interpretation of Kappa values (18).

Unless otherwise stated, SAS software was used for all statistical analyses (Statistical Analysis System, version 9.4, SAS Institute).

Results

Description of the population

Nine hundred seventy-two subjects were initially recruited, of whom 52 were excluded because of abnormal values in the standard laboratory screening tests. A further nine subjects were excluded because of missing information on pregnancy status or viral serology. The study population thus consisted of 911 subjects (470 males), comprising 101, 118, 99, 98, 103, 102, 108, 97, and 85 subjects in the 18–20, 21–23, 24–26, 27–29, 30–39, 40–

Downloaded from https://academic.oup.com/jcem/article/101/9/3450/2806775 by guest on 04 May 2025

Table 2. Normative Reference Intervals (95% CI) of IGF-I Measured by Six Assay Methods According to Age Range and Sex in a Cohort of 899 Healthy Subjects

Age Range	N	iSYS IGF-I (ng/ml) 95% CI	LIAISON XL IGF-I (ng/ml) 95% CI	IMMULITE 2000 IGF-I (ng/ml) 95% CI	IGFI-RIACT IGF-I (ng/ml) 95% CI	Mediagnost ELISA IGF-I (ng/ml) 95% CI	Mediagnost RIA IGF-I (ng/ml) 95% CI
Males (y)							
18–20	56	168–391	186–453	195–537	197–486	177–430	168–374
21–23	61	147–346	168–411	171–477	173–430	159–388	150–337
24–26	53	132–313	153–377	152–430	155–389	144–355	135–308
27–29	49	122–292	142–351	138–396	143–363	133–331	126–289
30–39	56	108–265	124–310	118–348	127–329	115–295	112–265
40–49	51	91–233	106–271	98–301	107–286	98–261	97–237
50–59	54	81–214	97–252	85–273	94–262	88–245	86–218
60–69	49	75–208	92–245	77–260	87–250	80–237	82–214
70–89	34	64–192	80–220	66–242	75–231	71–233	72–200
Females (y)							
18–20	41	155–421	191–483	180–586	169—517	169—487	161—412
21–23	54	144–383	176–448	166–541	159–476	156–446	149–379
24–26	45	134–353	163–418	153–501	150–440	144–412	139–353
27–29	48	126–330	152–391	142–467	142–410	134–385	131–332
30–39	47	113–294	131–345	121–403	126–356	118–341	118–298
40–49	50	97–253	109–296	98–331	107–297	100–296	103–258
50–59	54	80–209	93–253	80–271	90–247	82–248	97–220
60–69	47	64–170	84–222	68–227	76–209	68–208	75–190
70–89	50	56–154	81–204	60–188	67–189	60–187	68–175

49, 50–59, 60–69, and 70–89 year age groups, respectively. Mean BMI was 23.0 ± 2.4 kg/m^2.

IGF-I reference intervals obtained with the six assays

The IGF-I reference intervals (2.5th–97.5th percentiles) obtained with the six immunoassays are shown in Table 2 according to age and sex. Supplemental Figure 1 shows individual points and fitted percentiles (2.5%, 50%, and 97.5%) for males and females in each IGF-I assay.

A calculator available online (http://ticemed_sa.upmc.fr/sd_score/) or by using Apps (IGF1 SD_score) downloadable for Android from Google Play and for iOS from Apple Store (free of charge) allows the obtaining of individual IGF-I SDS after entering the name of the assay, the individual IGF-I value obtained with the assay, and the sex and age of the individual.

The six reference intervals for males and females are plotted on the same graph in Figure 1. Although the lower limits of the reference intervals (2.5th percentiles) were similar, the upper limits (97.5th percentiles) varied markedly from one assay to another.

3-Comparison of IGF-I levels given by the six assays

The results obtained with each IGF-I assay were compared with those obtained with each of the other five assays. Scatter plots and Bland-Altman plots based on raw values and SDS for each pair of assays are shown in Supplemental Figure 2.

Whatever the assay, IGF-I concentrations were generally higher in women than in men until the age of 59 years (this was significant for the age ranges 18–20 and 24–26 years). From the age of 60 years, IGF-I levels were slightly higher in men than in women, although the gender difference was smaller than in the younger age groups and was only significant for Immulite, Mediagnost ELISA, and Mediagnost RIA.

Two examples of interassay comparisons are shown in Figure 2. The results obtained with iSYS and Mediagnost RIA were in good overall agreement, with no significant bias as assessed by Bland-Altman plots (Figure 2, A–D). In contrast, the results obtained with LIAISON XL and Mediagnost RIA were not in good agreement (Figure 2, E–H).

Pairwise assay concordances assessed with the weighted Kappa coefficient for categorized IGF-1 SDS are shown in Table 3. The concordances were moderate to good (0.38–0.70), although the percentages of observed agreement were quite high (94–97%).

Overall agreement was moderate as overall Kappa coefficient was 0.55. Both in men and women, global interassay comparison showed significant differences ($P < .0001$) on raw values but not on SDS values ($P = .26$ and $P = .36$, respectively).

Table 4 shows pairwise concordances between the reference intervals provided by the manufacturer and those

Downloaded from https://academic.oup.com/jcem/article/101/9/3450/2806775 by guest on 04 May 2025

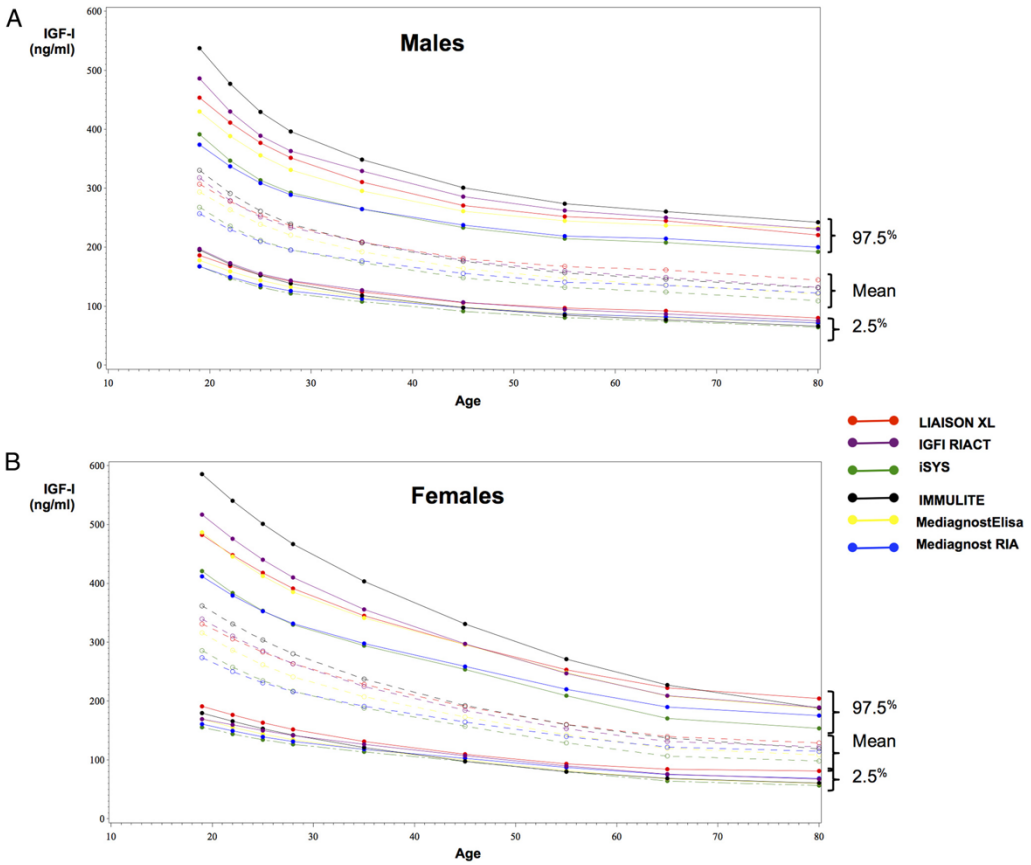

Figure 1. Reference intervals for (A) males and (B) females according to the age intervals of the six IGF-I immunoassays tested. Lower limits (2.5th percentile) and upper limits (97.5th percentile) of the normal range are drawn as full lines and means as dotted lines.

obtained in the VARIETE cohort, as assessed by the Kappa coefficient and the percentage agreement for each IGF-I assay. The concordances and percentages of observed agreement were generally poor.

Discussion

We report reference intervals for IGF-I concentrations obtained with six immunoassays in the same population of nearly 900 French healthy subjects aged 18–90 years, in keeping with the 2011 recommendations of the Consensus Group on the Standardization and Evaluation of GH and IGF-I assays (1). The population composed about 100 subjects per age decade, and specific reference intervals were calculated for each sex and age group. The reference intervals varied from one assay to another: the lower limits of the normal range (2.5th percentile) were quite similar with the six methods, but the upper limits (97.5th percentile) varied widely from one assay to another, in both men and women (Figure 1). Although the preanalytic conditions were the same for the six kits, and although four of the six kits were calibrated against the international reference standard 02/254, concordance between the assays, as assessed with Bland-Altman plots and the Kappa coefficient, remained quite variable, not only for raw IGF-I values but also for IGF-I SDS. This latter result was somewhat surprising, because we expected that, by using the same healthy population, we would obtain similar SDS.

In Table 2, which shows the reference ranges for each assay, we have deliberately omitted the mean and SD calculated for each age category from the raw values to avoid erroneous calculations of SDS. Indeed, the Box-Cox power transformation, which is necessary because of the non-Gaussian distribution in each age category, uses pa-

Downloaded from https://academic.oup.com/jcem/article/101/9/3450/2806775 by guest on 04 May 2025

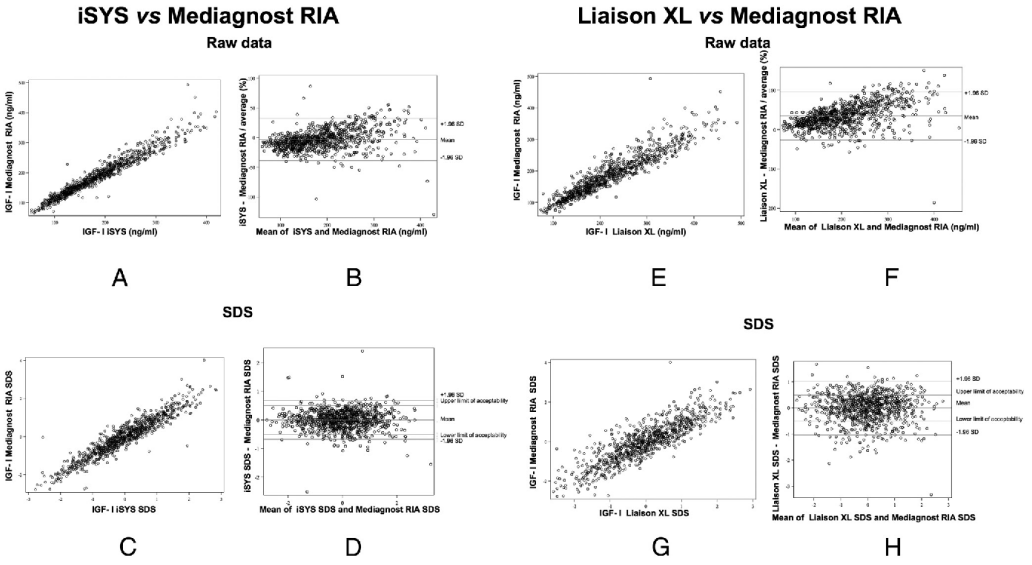

Figure 2. Comparisons between iSYS and Mediagnost RIA expressed as scatter plots (A) or Bland-Altman plots (B) for raw data, or scatter plots (C) and Bland-Altman plots (D) for SDS showing a good overall agreement between both IGF-I immunoassays, with no significant bias. Comparisons between Liaison XL and Mediagnost RIA expressed as scatter plots (E) or Bland-Altman plots (F) for raw data, or scatter plots (G) and Bland-Altman plots (H) for SDS showing a bad overall agreement between these two immunoassays.

rameters (L, M, and S) that are specific to each assay and also to each age group and gender. We thus propose an online calculator available either following this link (http://ticemed_sa.upmc.fr/sd_score/) or by using Apps (IGF-I SD_score) downloadable for Android from Google Play and for iOS from Apple Store (free of charge), which allows the determination of SDS as a function of the assay method, the measured IGF-I value, gender, and age. L, M, and S parameters are also provided in Supplemental Table 1.

Reliable reference intervals are crucial for interpreting IGF-I values in adults with acromegaly (for diagnosis and assessment of disease control during treatment), and also for diagnosing GH deficiency and monitoring GH therapy (4, 5, 19, 20). Reference intervals obtained with the IGF-I Nichols Advantage assay in a very large population of healthy subjects (21) were once widely used for research and clinical practice. However, market withdrawal of this assay, together with the availability of numerous automated methods with considerable heterogeneity, led to calls for improved comparability and reliable normative data. One important first step was the creation of the recombinant international IGF-I standard preparation 02/254 (22). A consensus conference held in 2011 proposed

Table 3. Agreement of Each IGF-I Assay Method Against Each of the Others, Expressed as Weighted Kappa and Percentages of Observed Agreement

	LIAISON XL	iSYS	IMMULITE 2000	Mediagnost ELISA	Mediagnost RIA	IGFI-RIACT
LIAISON XL	—	0.49 94.86%	0.50 94.83%	0.47 94.95%	0.38 94.05%	0.48 95.22%
iSYS	0.49 94.86%	—	0.64 96.08%	0.61 96.11%	0.70 97.00%	0.64 96.46%
IMMULITE 2000	0.50 94.83%	0.64 96.08%	—	0.61 95.95%	0.58 95.73%	0.64 96.32%
Mediagnost ELISA	0.47 94.95%	0.61 96.11%	0.61 95.95%	—	0.59 96.00%	0.53 95.66%
Mediagnost RIA	0.38 94.05%	0.70 97.00%	0.58 95.73%	0.59 96.00%	—	0.48 95.22%
IGFI- RIACT	0.48 95.22%	0.64 96.46%	0.64 96.32%	0.53 95.66%	0.48 95.22%	—

Downloaded from https://academic.oup.com/jcem/article/101/9/3450/2806775 by guest on 04 May 2025

Table 4. Concordance Between IGF-I VARIETE Cohort Reference Intervals and IGF-I Reference Intervals Provided by Each Manufacturer, Expressed as Kappa and Percentages of Observed Agreement

	LIAISON XL	iSYS	IMMULITE 2000	Mediagnost ELISA	Mediagnost RIA	IGFI -RIACT
Weighted Kappa	0.19	0.35	0.38	0.18	0.17	0.22
% of agreement	83.28	93.36	86.97	93.55	94.77	88.21

that all assays be calibrated against this standard, and advocated precise preanalytical and analytical conditions (1). Another recommendation was to establish normative data based on a random selection of individuals from the background population, with representation of all age groups (1). The first normative data for the iSYS IGF-I assay, based on these recommendations and on a very large healthy population, were published by Bidlingmaier et al (23). We now propose reference intervals for six IGF-I assays also based on a large population of healthy subjects. It should be noted that we used very stringent inclusion criteria. Indeed, despite the large sample size (almost 1000 healthy subjects, with about 100 subjects per age group), all the subjects had a clinical examination, including assessment of gonadal status, and also a careful medical history-taking that included ongoing medications. Furthermore, all the subjects had an extensive standard biological workup to exclude those with disorders capable of influencing IGF-I levels or their measurement. These very strict inclusion and exclusion criteria allow us to define a population as "healthy" as possible; however, this implies that these normative data will not be strictly applicable to patients with BMI higher than 28 kg/m^2 or to patients with oral treatment with estrogens.

As expected, IGF-I concentrations fell gradually with age in both sexes, irrespective of the assay. Contrary to previous reports (21, 23), we found a gender difference, with higher IGF-I levels in women than in men, whatever the assay, until the fifth decade. After 50 years of age, however, IGF-I levels were higher in men than in women, as reported elsewhere (21, 23). We therefore propose separate normative data for men and women. One possible explanation for the discrepancy between this work and previous reports is that we excluded all subjects receiving steroid hormones such as estrogens. Indeed, oral estrogen is known to lower IGF-1 levels (6–8). In premenopausal women, for example, contraceptive pills containing ethinyl estradiol reduce IGF-I levels by up to an average of 30% (24–27). Another explanation might be the size of our population. Indeed, in their study involving a larger number of subjects (15 000), Bidlingmaier et al did not find differences in terms of gender differences (23).

Interassay differences in IGF-I reference intervals are a well-known issue that has previously been underlined by one of us (28, 29) and by many other researchers (2, 11, 23, 30, 31). In this study, as expected, the largest intercentile intervals (and highest values) were obtained with the two assays calibrated with the old standard IRP 87/518 (IMMULITE and IGFI RIACT). Moreover, the three automated methods (iSYS, Liaison XL, and IMMULITE), which should theoretically be the most reproducible, did not yield narrower reference intervals. For example, the iSYS automated method and the Mediagnost RIA manual method gave very similar intervals for both men and women in all age groups. Thus, the main source of variation does not appear to be analytical reproducibility. Using the same iSYS method and a similar transformation for normalizing data and constructing specific centile curves in the LMS method, our 2.5th and 97.5th percentiles were generally slightly higher and our intervals generally narrower than those reported by Bidlingmaier et al (23). Although interlaboratory variability may play a role in these discrepancies, they are likely due mainly to differences in the population samples (our population was smaller, and the inclusion criteria were different). Another issue raised by our study is the poor concordance between our reference intervals and those proposed by the assay manufacturers. Once again, this might be related to the use of different background populations: indeed, those used by manufacturers may not fulfill all the criteria recommended by the consensus group in 2011, particularly with respect to their size, the definition of healthy subjects, and the use of hormonal contraceptives (Supplemental Material).

Likewise, one obvious explanation for the discordance between assays is the use of different populations to establish reference intervals. This is why we used the same reference population for all the kits. However, although the six assays showed comparable analytical performance in terms of their reproducibility and detection limits (Table 1), and despite the fact that they use the same noncompetitive "sandwich" format and similar methods to avoid IGFBP interference (IGF-II addition), the reference values obtained in our well-controlled adult population differed strikingly from one assay to another. Two of the six assays (IMMULITE and IGF-I RIACT) are still calibrated against the old international reference reagent IRR 87/518 standard, whereas the other four are calibrated against the new IRR 02/254 standard, as currently recommended (1). As expected, the former two assays gave the highest upper reference range for both sexes until the

Downloaded from https://academic.oup.com/jcem/article/101/9/3450/2806775 by guest on 04 May 2025

age of 50 years (Table 2, Figure 1). However, the reference ranges of two differently calibrated kits may be either similar (eg, LIAISON XL and IGFI RIACT in men) or significantly different (eg, iSYS lower than IMMULITE) (Table 2). Likewise, reference ranges determined with kits calibrated against the same reference preparation may also be significantly different, even for kits from the same manufacturer (eg, the RIA and ELISA kits from Mediagnost). It therefore seems likely that the observed differences are related to other analytical factors, such as the efficiency of IGFBP interference removal and the specificity and/or affinity of the antibody used. For example, since the 2.5th percentile is at least similar between the assays, the broadening of the interval for the IMMULITE assay is probably not related to the calibrator, but to relatively higher measurement results at the upper end: an explanation could be that IMMULITE assay preferentially recognizes the high free IGF-I at high concentrations, whereas the other two assays more efficiently remove the impact of binding proteins.

This could have important implications in patients with disorders affecting their IGFBP profile, such as acromegaly and chronic kidney disease. If confirmed in further studies, this implies that a given individual must be monitored with the same IGF-I assay.

Another limitation of our study is that it lies on a single measurement of IGF-I while it is well known that there is some within-subject variability when an individual is sampled on different days (32, 33).

What refinements may be expected in the measurement of this very demanding analyte? The liquid chromatography (LC) tandem mass spectometry (MS) method may prove to be a valid alternative and is now being used to assess interlaboratory agreement on IGF-I concentrations (34) or for validation of IGF-I measures (35). Reference intervals for IGF-I provided with this LC-MS (36) seem very comparable with those obtained with immunoassays. When compared with our data, the lower limit of normal range is similar and upper limit corresponds more or less with those observed with Liaison XL or IGF1 RIACT immunoassays. However, tandem LC-MS is a time-consuming and complex method that requires expensive machines and high technical expertise, because many variables need to be controlled for providing accurate quantitative results (eg, extraction strategies, approaches to detect and quantify IGF-I, calibration protocols) (37). Furthermore, a recent preliminary study of an LC-MS method suggested that it might miss some IGF-I protein variants (pathogenic or physiological), which are present in 0.6% of the population (38). Thus, despite their limitations, immunoassays will continue to be widely used, at least in the near future (39).

In conclusion, we have established reference intervals for six commercial IGF-I assays in a study conforming to recent international recommendations. Despite being obtained in the same large population of French healthy subjects, the reference intervals differed somewhat from one assay to another, and agreement between assays was moderate to good. Finally, concordances between the manufacturers' reference intervals and those obtained in our cohort were generally poor. These findings confirm the need to establish reference intervals for each commercial IGF-I assay in a large background population. Interassay concordance with respect to the classification of patients with acromegaly or GH deficiency remains to be determined, and the IGF-I standard deviation scores obtained with the six assays in these subjects need to be compared.

Acknowledgments

The authors thank Dr. Hélène Agostini and Prof. Bruno Falissard for their helpful comments. The authors also thank Cisbio International, DiaSorin, IDS, Mediagnost, and Siemens for the kind donation of IGF-I kits.

Address all correspondence and requests for reprints to: Philippe Chanson, MD, Service d'Endocrinologie et des Maladies de la Reproduction, Hôpital de Bicêtre, 78 rue du Général Leclerc, 94275 Le Kremlin-Bicêtre, France. E-mail: philippe.chanson@bct.aphp.fr.

This study was supported by a grant from Programme Hospitalier de Recherche Clinique, French Ministry of Health, no. P081216/IDRCB 2009-A00892-55, to Drs Chanson and Souberbielle, and by a grant from Fond National Suisse, P1GEP3 155694, to Dr. Mavromati.

*A complete list of the VAleurs de Référence de l'IGF-I Et Transformation En Z-score (VARIETE) study investigators is given in Supplemental Appendix.

Trial Registration: ClinicalTrials.gov: NCT01831648.

Disclosure Summary: The authors have nothing to disclose.

References

1. Clemmons DR. Consensus statement on the standardization and evaluation of growth hormone and insulin-like growth factor assays. *Clin Chem*. 2011;57:555–559.
2. Pokrajac A, Wark G, Ellis AR, Wear J, Wieringa GE, Trainer PJ. Variation in GH and IGF-I assays limits the applicability of international consensus criteria to local practice. *Clin Endocrinol (Oxf)*. 2007;67:65–70.
3. Quarmby V, Quan C, Ling V, Compton P, Canova-Davis E. How much insulin-like growth factor I (IGF-I) circulates? Impact of standardization on IGF-I assay accuracy. *J Clin Endocrinol Metab*. 1998;83:1211–1216.
4. Frystyk J, Freda P, Clemmons DR. The current status of IGF-I assays–a 2009 update. *Growth Horm IGF Res*. 2010;20:8–18.
5. Brabant G, Wallaschofski H. Normal levels of serum IGF-I: determinants and validity of current reference ranges. *Pituitary*. 2007; 10:129–133.

Downloaded from https://academic.oup.com/jcem/article/101/9/3450/2806775 by guest on 04 May 2025

6. Juul A. Serum levels of insulin-like growth factor I and its binding proteins in health and disease. *Growth Horm IGF Res*. 2003;13:113–170.
7. Leung KC, Johannsson G, Leong GM, Ho KK. Estrogen regulation of growth hormone action. *Endocr Rev*. 2004;25:693–721.
8. Meinhardt UJ, Ho KK. Modulation of growth hormone action by sex steroids. *Clin Endocrinol (Oxf)*. 2006;65:413–422.
9. Clemmons DR. Value of insulin-like growth factor system markers in the assessment of growth hormone status. *Endocrinol Metab Clin North Am*. 2007;36:109–129.
10. Massart C, Poirier JY, Jard C, Pouchard M, Vigier MP. Determination of serum insulin-like growth factor-I reference values for the immunometric Cisbio method on a large number of healthy subjects: clinical utility in the follow-up of patients with treated acromegaly. *Clin Chim Acta*. 2007;381:176–178.
11. Varewijck AJ, Lamberts SW, van der Lely AJ, Neggers SJ, Hofland LJ, Janssen JA. The introduction of the IDS-iSYS total IGF-1 assay may have far-reaching consequences for diagnosis and treatment of GH deficiency. *J Clin Endocrinol Metab*. 2015;100:309–316.
12. Cole TJ, Green PJ. Smoothing reference centile curves: the LMS method and penalized likelihood. *Stat Med*. 1992;11:1305–1319.
13. Blum WF, Breier BH. Radioimmunoassays for IGFs and IGFBPs. *Growth Regul*. 1994;4 Suppl 1:11–19.
14. Royston P, Wright EM. A method for estimating age-specific reference intervals based on fractional polynomials and exponential transformation. *J Royal Stat Soc Series A*. 1998;161:79–101.
15. Rigby RA, Stasinopoulos DM. Generalized additive models for location, scale and shape. *J Royal Stat Soc Series C (Applied Statistics)*. 2005;54:507–554.
16. Fleiss J-L, Levin B, Paik MC. *Statistical Methods for Rates and Proportions*. 3rd ed. New York: Wiley.
17. Cicchetti DV, Allison T. A new procedure for assessing reliability of scoring EEG sleep recordings. *Am J EEG Technol*. 1971;11:101–109.
18. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159–174.
19. Clemmons DR. IGF-I assays: current assay methodologies and their limitations. *Pituitary*. 2007;10:121–128.
20. Junnila RK, Strasburger CJ, Bidlingmaier M. Pitfalls of insulin-like growth factor-I and growth hormone assays. *Endocrinol Metab Clin North Am*. 2015;44:27–34.
21. Brabant G, von zur Muhlen A, Wuster C, et al. Serum insulin-like growth factor I reference values for an automated chemiluminescence immunoassay system: results from a multicenter study. *Horm Res*. 2003;60:53–60.
22. Burns C, Rigsby P, Moore M, Rafferty B. The First International Standard For Insulin-like Growth Factor-1 (IGF-1) for immunoassay: preparation and calibration in an international collaborative study. *Growth Horm IGF Res*. 2009;19:457–462.
23. Bidlingmaier M, Friedrich N, Emeny RT, et al. Reference intervals for insulin-like growth factor-1 (IGF-I) from birth to senescence: results from a multicenter study using a new automated chemiluminescence IGF-I immunoassay conforming to recent international recommendations. *J Clin Endocrinol Metab*. 2014;99:1712–1721.
24. Jernstrom H, Deal C, Wilkin F, et al. Genetic and nongenetic factors associated with variation of plasma levels of insulin-like growth factor-I and insulin-like growth factor-binding protein-3 in healthy premenopausal women. *Cancer Epidemiol Biomarkers Prev*. 2001;10:377–384.
25. Elkazaz AY, Salama K. The effect of oral contraceptive different patterns of use on circulating IGF-1 and bone mineral density in healthy premenopausal women. *Endocrine*. 2015;48:272–278.
26. Blackmore KM, Wong J, Knight JA. A cross-sectional study of different patterns of oral contraceptive use among premenopausal women and circulating IGF-1: implications for disease risk. *BMC Womens Health*. 2011;11:15.
27. Balogh A, Kauf E, Vollanth R, et al. Effects of two oral contraceptives on plasma levels of insulin-like growth factor I (IGF-I) and growth hormone (hGH). *Contraception*. 2000;62:259–269.
28. Massart C, Poirier JY. Serum insulin-like growth factor-I measurement in the follow-up of treated acromegaly: comparison of four immunoassays. *Clin Chim Acta*. 2006;373:176–179.
29. Massart C, Poirier JY. Determination of serum insulin-like growth factor-I reference values for the automated chemiluminescent Liaison(R) assay. Clinical utility in the follow-up of patients with treated acromegaly. *Clin Chim Acta*. 2011;412:398–399.
30. Krebs A, Wallaschofski H, Spilcke-Liss E, et al. Five commercially available insulin-like growth factor I (IGF-I) assays in comparison to the former Nichols Advantage IGF-I in a growth hormone treated population. *Clin Chem Lab Med*. 2008;46:1776–1783.
31. Cowan DA, Bartlett C. Laboratory issues in the implementation of the marker method. *Growth Horm IGF Res*. 2009;19:357–360.
32. Milani D, Carmichael JD, Welkowitz J, et al. Variability and reliability of single serum IGF-I measurements: impact on determining predictability of risk ratios in disease development. *J Clin Endocrinol Metab*. 2004;89:2271–2274.
33. Nguyen TV, Nelson AE, Howe CJ, et al. Within-subject variability and analytic imprecision of insulinlike growth factor axis and collagen markers: implications for clinical diagnosis and doping tests. *Clin Chem*. 2008;54:1268–1276.
34. Cox HD, Lopes F, Woldemariam GA, et al. Interlaboratory agreement of insulin-like growth factor 1 concentrations measured by mass spectrometry. *Clin Chem*. 2014;60:541–548.
35. Kay R, Halsall DJ, Annamalai AK, et al. A novel mass spectrometry-based method for determining insulin-like growth factor 1: assessment in a cohort of subjects with newly diagnosed acromegaly. *Clin Endocrinol (Oxf)*. 2013;78:424–430.
36. Bystrom C, Sheng S, Zhang K, et al. Clinical utility of insulin-like growth factor 1 and 2; determination by high resolution mass spectrometry. *PLoS One*. 2012;7:e43457.
37. Hoofnagle AN, Whiteaker JR, Carr SA, et al. Recommendations for the generation, quantification, storage, and handling of peptides used for mass spectrometry-based assays. *Clin Chem*. 2016;62:48–69.
38. Hines J, Milosevic D, Ketha H, et al. Detection of IGF-1 protein variants by use of LC-MS with high-resolution accurate mass in routine clinical analysis. *Clin Chem*. 2015;61:990–991.
39. Ketha H, Singh RJ. Clinical assays for quantitation of insulin-like-growth-factor-1 (IGF1). *Methods*. 2015;81:93–98.

Downloaded from https://academic.oup.com/jcem/article/101/9/3450/2806775 by guest on 04 May 2025

iii. Classification of patients with GH disorders may vary according to the IGF-1 assay

Diagnosis and follow-up of GH disorders (acromegaly, GH-deficiency) relies on accurate IGF-1 measurement. However, commercially available assays provide different IGF-1 results as well as different patient classification to normal, low or high IGF-1 levels, which further complicates management. This variability in IGF-1 values among different assays persists even if the same international calibrator, suggested by the WHO consensus, is used, and mainly results from the fact that IGF-1 varies with age, sex and BMI and from the differences among reference populations used to establish normative data. The VARIETE Study, previously described aimed to establish age- and sex-specific, normative IGF-1 values, for 6 widely used IGF-1 immunoassays (iSYS, Liaison XL, Immulite 2000, IGF-1 RIACT, Mediagnost ELISA, Mediagnost RIA), using a well-chosen healthy adult population. The study provided normal reference values for each one of these assays as well as a formula for SD score calculation, yet, showed moderate to good agreement between methods.

The reference normative data issued from the VARIETE study were further tested for the classification of patients with GH disorders in this cross-sectional study. Altogether 102 patients were included (56 patients with acromegaly, 14 patients with GH deficiency and 32 patients with suspected GH disorder). For each individual patient, IGF-1 levels were measured at a specific time point (either upon diagnosis or during follow-up) with the 6 assays included in the VARIETE study, while 6 patients had several IGF-1 measurements. Pairwise comparisons between assays were made both for raw data and SD scores and were illustrated with scatter plots and Bland-Altman plots. Pairwise concordances were also calculated by means of weighted κ coefficient; the best concordance was between Mediagnost RIA and iSYS with a 0.81 κ coefficient (excellent agreement) and the worst concordance between iSYS and Liaison XL with a 0.50 κ coefficient (moderate agreement). Concordance between assays was better for low and high IGF-1 values, while for normal IGF-1 values, concordance was moderate to poor.

In this study, since the reference normative values for IGF-1 were obtained from the same healthy population after careful exclusion of confounding factors and given that the preanalytical procedure was identical, it is surprising that concordance among immunoassays remains variable. This can be explained by the differences in epitope specificity as well as the differences in the elimination of IGF-binding proteins, the latter being particularly important in patients with acromegaly, associated with higher levels of IGF-BP3.

In conclusion, even if reference values of IGF-1 provided by the same large healthy population are used, commercially available IGF-1 immunoassays show variable agreement for the classification of patients with GH disorders, especially for values close to the normal range. Performance characteristics of each immunoassay, such as the tendency to give higher or lower IGF-1 values than other assays are an important knowledge for the clinician. The follow-up of each patient needs to be performed with the same immunoassay, or at least with assays that share similar characteristics.

Classification of Patients With GH Disorders May Vary According to the IGF-I Assay

Maria Mavromati,[1] Emmanuelle Kuhn,[1,2] Hélène Agostini,[3] Sylvie Brailly-Tabard,[2,4] Catherine Massart,[5] Marie-Liesse Piketty,[6] Armelle Arnoux,[3] Jacques Young,[1,2] Jean-Claude Souberbielle,[6] and Philippe Chanson[1,2]

[1]Service d'Endocrinologie et des Maladies de la Reproduction and Centre de Référence des Maladies Rares de l'Hypophyse, F94275 Le Kremlin-Bicêtre, France; [2]Inserm 1185, Fac Med Paris Sud, Université Paris-Saclay, F94276 Le Kremlin-Bicêtre, France; [3]Assistance Publique-Hôpitaux de Paris, Hôpitaux Universitaires Paris-Sud, Hôpital de Bicêtre, Unité de Recherche Clinique, F94275 Le Kremlin-Bicêtre, France; [4]Service de Génétique Moléculaire, Pharmacogénétique et Hormonologie, F94275 Le Kremlin-Bicêtre, France; [5]Laboratoire d'Hormonologie, Centre Hospitalier Universitaire de Rennes, Centre d'Investigation Clinique Plurithématique, Inserm 1414, Hôpital Pontchaillou, F29000 Rennes, France; and [6]Service des Explorations Fonctionnelles, Assistance Publique-Hôpitaux de Paris, Hôpital Necker-Enfants Malades, F75015 Paris, France

Context: Insulinlike growth factor I (IGF-I) measurement is essential for the diagnosis and management of growth hormone (GH) disorders. However, patient classification may vary substantially according to the assay technique.

Objective: We compared individual patient data and classifications obtained with six different IGF-I assay kits in a group of patients with various GH disorders.

Design: In this cross-sectional study, we measured IGF-I with six immunoassays in 102 patients with active or treated acromegaly or GH deficiency. IGF-I normative data previously established for the same six assay kits were used to classify the patients (high, low, or normal IGF-I levels), using both raw data and standard deviation scores (SDSs). Pairwise concordance between assays was assessed with Bland-Altman plots and with the percentage of observed agreement and the weighted κ coefficient for categorized IGF-I SDS.

Results: We observed marked variability both across each individual's IGF-I raw data and across IGF-I SDS values obtained with each of the six immunoassays. Pairwise concordance between assay values, as assessed with the weighted κ coefficient, ranged from 0.50 (moderate) to 0.81 (excellent).

Conclusion: Even when using normative data obtained in the same large population of healthy subjects and when using calculated IGF-I SDSs, agreement among IGF-I assay methods is only moderate to good. Differences in assay performance must be taken into account when evaluating and monitoring patients with GH disorders. This argues for the use of the same IGF-I assay for a given patient throughout follow-up. (*J Clin Endocrinol Metab* 102: 2844–2852, 2017)

Downloaded from https://academic.oup.com/jcem/article/102/8/2844/3819477 by guest on 01 June 2025

Insulinlike growth factor I (IGF-I) measurement is of crucial importance for the diagnosis of acromegaly and growth hormone deficiency (GHD), as well as for treatment monitoring (1). The Endocrine Society clinical practice guidelines for acromegaly, and the Acromegaly Consensus Group, recommend IGF-I measurements rather than random growth hormone (GH) values for diagnosis and treatment goals (2, 3). In patients with GHD, IGF-I is also crucial for monitoring GH replacement therapy and for adjusting the GH dosage (4).

ISSN Print 0021-972X ISSN Online 1945-7197
Printed in USA
Copyright © 2017 Endocrine Society
Received 21 January 2017. Accepted 8 May 2017.
First Published Online 12 May 2017

Abbreviations: GH, growth hormone; GHD, growth hormone deficiency; IGF-I, insulinlike growth factor I; LC-MS, tandem liquid chromatography and mass spectrometry; SDS, standard deviation score.

Accurate measurement of IGF-I is a complex issue, as the results depend on the type of analytical method. This variability can be attributed to differences in the calibration material, the epitope specificity of the different antibodies, and interference with IGF-I binding proteins (1, 5). A universal calibrator is crucial for assay standardization. A recent consensus statement on the evaluation and standardization of IGF-I assays recommends the IS 02/254 World Health Organization reference standard, a >97%-pure recombinant standard that has been well characterized by the National Institute for Biological Standards and Control (6).

Even if they give different results, one would expect two different IGF-I assays to classify a given patient in the same way (high, normal, or low values). However, even when using kits that are calibrated against the same international standard, and the same method to remove IGF-I binding proteins, patient classification in terms of IGF-I categories remains variable (7–10). We suspected that a potential reason for these discrepancies was the use of different reference values. Indeed, it is difficult to establish IGF-I normative data, as they depend on the choice of a healthy reference population (6, 11, 12). Although IGF-I values depend on many factors, such as sex, age, nutritional status, treatments (especially hormonal medications), diabetes, and renal and hepatic failure, normative data used for the different IGF-I kits were not obtained in the same, apparently healthy, population. Furthermore, the distribution of IGF-I levels in healthy populations is non-Gaussian, and transformations are thus necessary to obtain normal distributions and to calculate standard deviation scores (SDSs). This prompted us to conduct the VARIETE study (VAleurs de Référence de l'IGF-I Et Transformation En z score) to establish normative reference values for six IGF-I immunoassays in the same healthy adult population, using the same statistical method to calculate SDSs (13). We postulated that this would help to longitudinally assess disease control in a given patient using the IGF-I SDSs, even if IGF-I was measured with more than one assay during follow-up.

In the current study, we measured IGF-I with the same six kits in 102 patients with acromegaly or GH deficiency, and used the age- and sex-adjusted normative reference values from the VARIETE study to compare the raw data and SDS values obtained for each patient with each assay. We thus determined whether the patients' classifications were concordant.

Subjects and Methods

Study population

One hundred two patients (57 men and 45 women) belonging to the cohort of Service d'Endocrinologie et des Maladies de la Reproduction of Hôpitaux Universitaires Paris-Sud (Bicetre Hospital), Le Kremlin-Bicêtre, France, were enrolled in the study between December 2013 and March 2014. Fifty-six patients had acromegaly. Thirty-two patients had a blood sample taken at diagnosis (n = 11) or after incomplete surgery and before initiation of medical treatment (n = 21), and 24 patients were sampled during follow-up on medical treatment (cabergoline alone, n = 1; somatostatin analog alone, n = 10; pegvisomant alone, n = 9; somatostatin analog and cabergoline, n = 3; and somatostatin analog and pegvisomant, n = 1) but with variable disease control (because of treatment modification, reinforcement, or titration, or because they were resistant to medical treatment). Diagnosis of acromegaly was based on clinical criteria, unsuppressed GH in the oral glucose tolerance test, IGF-I elevation, and imaging or histologic proof of a somatotroph pituitary adenoma after surgery (2, 14, 15). Fourteen patients had GHD, either confirmed by a serum GH peak less than 5 μg/L after the insulin tolerance test (six patients) or strongly suggested by deficiencies in at least three other pituitary functions (4). Another 32 patients had other pituitary or endocrine disorders and were tested for suspected acromegaly or GHD. The patients' characteristics are summarized in Table 1. Each patient underwent a clinical examination, had a personal medical history obtained and was sampled at 8:00 AM after an overnight fast. Six patients had serial IGF-I measurements with three to six IGF-I assays (at diagnosis, after pituitary surgery, and during medical treatment with somatostatin analogs). All the patients gave their written informed consent to participate in the study, which was approved by the Paris-Sud Ethics committee.

In each patient, IGF-I values were measured with the six assay kits (see later) used in the recently published VARIETE study (13). The main characteristics of the assays are shown in Supplemental Table 1.

Normative reference range

The normative reference data that we used to classify patients as having "normal," "high," or "low" IGF-I levels were obtained in the VARIETE study (13). In brief, this study was a cross-sectional, multicenter (24 centers), nationwide French cohort study (ClinicalTrials.gov no. NCT01831648) designed to develop reference normative sex- and age-adjusted IGF-I data for the adult general population for each of the different assay techniques widely used in everyday clinical practice in France. The objective of this study was also to propose formulas for calculating IGF-I SDSs, taking into account the non-normal distribution of IGF-I levels in the healthy population. The study population consisted of 911 subjects (470 males), comprising 101, 118, 99, 98, 103, 102, 108, 97, and 85 subjects in the 18- to 20-, 21- to 23-, 24- to 26-, 27- to 29-, 30- to 39-, 40- to 49-, 50- to 59-, 60- to 69-, and 70- to 89-year age groups, respectively. Serum IGF-I was measured with the following six assay kits: iSYS (ImmunoDiagnostic Systems, Boldon, United Kingdom), Liaison XL (Diasorin, Saluggia, Italy), Immulite 2000 (Siemens, Erlangen, Germany), IGF-I RIACT (CIS BIO, Gif sur Yvette, France), Mediagnost ELISA, and Mediagnost RIA (Mediagnost, Reutlingen, Germany). IGF-I values were then matched in 3-year groups between 18 and 30 years of age and 10-year groups between 30 and 90 years, and mean and median values as well as the 2.5th and 97.5th percentiles were calculated. For each sex and age category, the distribution of measurements was normalized by means of sex- and age-specific

Downloaded from https://academic.oup.com/jcem/article/102/8/2844/3819477 by guest on 01 June 2025

Table 1. Characteristics of the 102 Patients With Various GH Disorders in Whom IGF-I Was Measured With the Six IGF-I Immunoassays

Characteristics	Males (n = 57)	Females (n = 45)
Age, y	47.1 (19–72)	43 (24–78)
Acromegaly (n = 56)		
Treated , n	11	13
Untreated, n	17	15
GHD (n = 14)		
GH-treated, n	0	1
Untreated, n	11	2
Suspicion of GH disorder (N = 32), n	18	14

Cox-Box power transformation to calculate SDSs. As men and women had significantly different IGF-I levels, curves were constructed separately using the LMS method.

The VARIETE study thus established age- and sex-specific adult normative data for the six commercial IGF-I assays, including the range of values from the 2.5th to the 97.5th percentile in mass units, and provided a formula for calculating SDSs. A calculator available online (http://ticemed_sa.upmc.fr/sd_score/) or as an app (IGF-I_SD_score) downloadable for Android from Google Play and for iOS from the Apple Store (free of charge) yields individual IGF-I SDSs after entering the name of the assay, the individual's IGF-I value obtained with the assay, and the sex and age of the individual.

Statistics

Data were analyzed with Statistical Analysis System software (version 9.4; SAS Institute, Cary, NC). We used scatter plots and Bland-Altman plots to assess pairwise concordance between assays, both for IGF-I raw values and SDS values. We classified the IGF-I results in three categories, high (SDS >+1.96), normal (SDS between −1.96 and +1.96), and low (SDS <−2), and evaluated pairwise agreement by means of the linearly weighted κ coefficient.

To interpret the κ coefficient, we used the Fermanian scale (16, 17), with κ values >0.80, between 0.61 and 0.80, between 0.41 and 0.60, between 0.21 and 0.40, between 0.01 and 0.20, and <0.01 signifying almost perfect, substantial, moderate, fair, slight, and poor agreement, respectively.

Results

Variability of individual IGF-I SDS values according to the IGF-I assay

Variability between each individual's IGF-I SDS obtained with each of the six immunoassays is illustrated in Fig. 1 for the 57 male patients and the 45 female patients. Many patients were inconsistently classified, particularly when IGF-I values were close to the reference range.

In six prospectively followed patients with acromegaly, IGF-I was measured on three occasions (at diagnosis, after surgery, and at follow-up, generally under medical treatment) with between three and six of the IGF-I assays (Fig. 2). With the exception of one patient in whom two of the three assays used at diagnosis gave a high IGF-I SDS, IGF-I SDSs were generally concordant in the elevated levels. In three out of six patients with borderline IGF-I SDS after surgery, IGF-I SDS was either normal using some assays, suggesting that the patient was cured, or moderately elevated using other assays, suggesting that the patient had persistent active disease. At follow-up under treatment, when IGF-I SDS was borderline, some assays classified the patient as "controlled," although others gave a low SDS.

Percentages of patients classified as having normal, high, or low IGF-I levels in the different IGF-I assays

The percentages of patients classified as having high (>+1.96), normal (between −1.96 and +1.96), and low (<−1.96) SDS values are shown in Fig. 3. The iSYS and Mediagnost RIA kits classified fewer patients as having "normal" levels (33% and 35%, respectively, *vs* 46% to 49% for the other assays) and, on the contrary, more patients as having "high" IGF-I values (54% and 51%, respectively, *vs* 30% to 39%). On the other hand, IGF-I RIACT and Liaison XL were more likely to classify the patients as having "low" IGF-I levels (20% and 23%, respectively, *vs* 13% to 16%).

Pairwise correlations between raw data and *z* scores obtained with the six IGF-I immunoassays

The results obtained with each IGF-I assay were compared with those obtained with each of the other five assays. Scatter plots and Bland-Altman plots based on raw values and SDSs for each pair of assays are shown in Supplemental Fig. 1.

Two examples of interassay comparisons are shown in Fig. 4. The results obtained with iSYS and Mediagnost RIA were in good overall agreement, with no significant bias on Bland-Altman plots [Fig. 4(a–d)]. Indeed, the discrepancy around the mean difference (average difference) line was quite stable when the average value increased, without very wide limits of agreement, and with consistent variability across the graph. In contrast, the results obtained with Liaison XL and iSYS were not in good agreement, as the mean difference line was clearly different from zero and as iSYS tended to overestimate IGF-I values by comparison with Liaison XL, an effect that was accentuated as the average value increased, especially for raw data [Fig. 4(e–h)].

Pairwise assay concordances (weighted κ coefficient) for categorized IGF-I SDS values are shown in Table 2. The best concordance was found between iSYS and Mediagnost RIA, with a κ coefficient of 0.81. Very good agreement was also observed between Immulite and Mediagnost ELISA (κ coefficient, 0.77), Mediagnost

Downloaded from https://academic.oup.com/jcem/article/102/8/2844/3819477 by guest on 01 June 2025

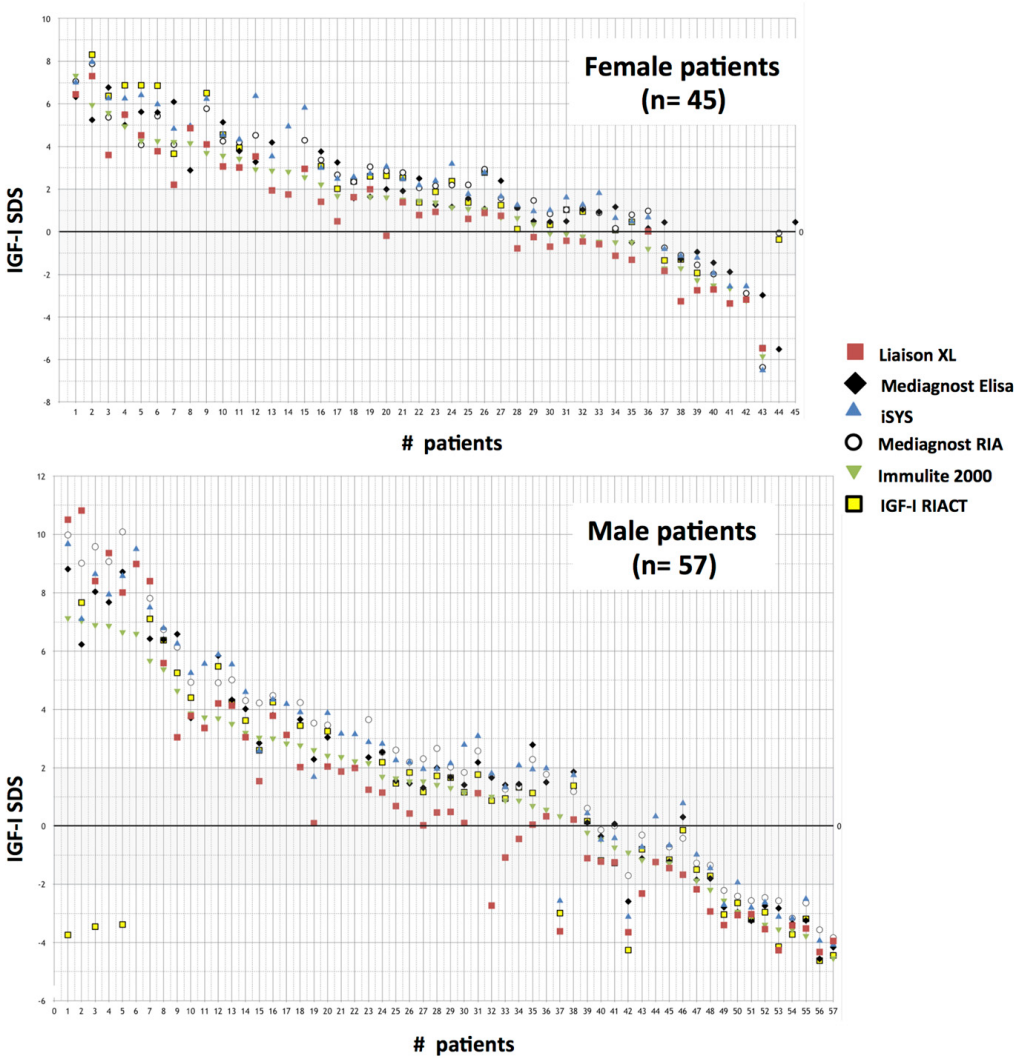

Figure 1. Variability among the six immunoassay SDS values for each of the 45 female patients (upper panel) and the 57 male patients (M; lower panel) with IGF-I disorder ranked by IGF-I Immulite 2000 decreasing value. Each assay is assigned a colored symbol.

ELISA and IGF-I RIACT (κ coefficient, 0.77), Immulite and Liaison XL (κ coefficient, 0.76), and Mediagnost ELISA and RIA (κ coefficient, 0.76), as well as between IGF-I RIACT and iSYS or Mediagnost RIA (κ coefficient, 0.71). The poorest concordance was observed between iSYS and Liaison XL (κ coefficient, 0.50), Mediagnost RIA and Liaison XL (κ coefficient, 0.51), and iSYS and Immulite (κ coefficient, 0.55) (Table 2).

When we limited the assessment of concordance to the group of patients with acromegaly (n = 56), the best pairwise agreement was again between iSYS and Mediagnost RIA, with a weighted κ coefficient of 0.81, whereas the worst agreement was between Liaison XL and iSYS and between Immulite and Mediagnost ELISA (κ coefficient, 0.41 for both).

Concordance between assays according to IGF-I SDS classes (high, normal, and low)

We analyzed concordance according to IGF-I SDS classes (high, >1.96; normal, between −1.96 and 1.96;

Downloaded from https://academic.oup.com/jcem/article/102/8/2844/3819477 by guest on 01 June 2025

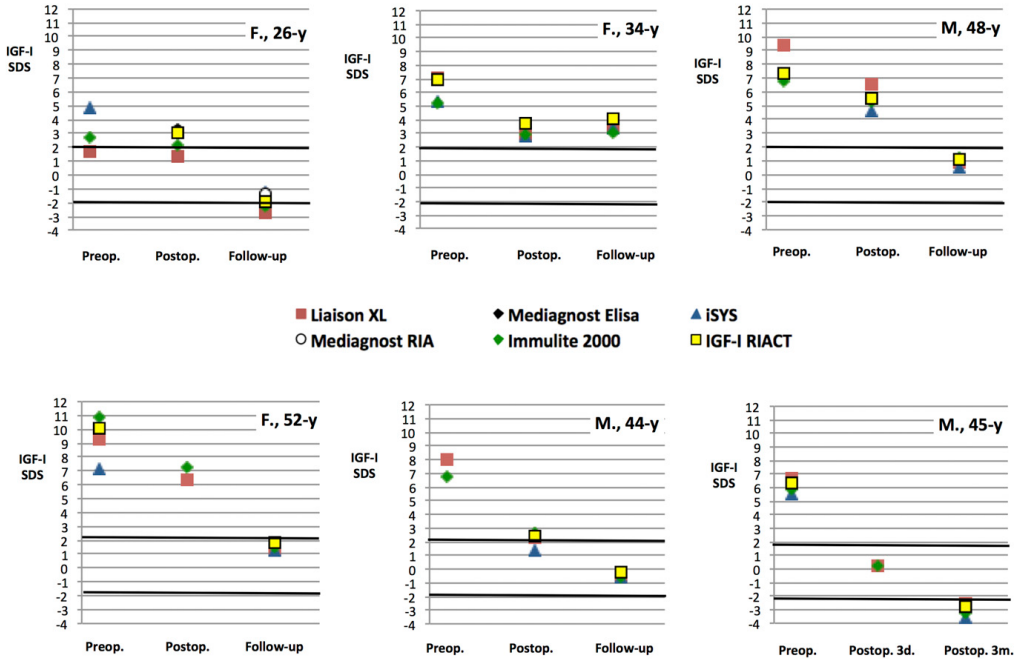

Figure 2. Variability of SDS values obtained with each of the six immunoassays in six patients with acromegaly, at three points of follow-up: diagnosis of acromegaly (Preop.), immediately after surgery (Postop. 3d), and at follow-up, generally under medical treatment. Horizontal lines represent the normal range from +2 to −2 standard deviations. Each assay is assigned a different colored symbol. Postop. 3m., 3 months after surgery.

and low, <−1.96). The three classes were those obtained initially with Immulite assay. Due to the small numbers, it was not always possible to calculate κ values for all comparisons. Thus we also calculated the concordance in terms of similar classification (as high, normal, or low values) between assays. The results are indicated as the ratio of concordant to total results in Supplemental Tables 2 to 7.

For low values (SDS < 1.96), assays are relatively concordant with only one or two patients (out of 11 to 14) who are discordantly classified by two assays (Liaison XL and iSYS, iSYS and Immulite, Liaison XL and Mediagnost ELISA, and Immulite and Mediagnost ELISA).

For "normal" IGF-I SDSs (between −1.96 and 1.96), concordances (as assessed by κ values) are generally weak or poor. In general, at least six patients out of around 40 are misclassified according to the assay that is used.

For high IGF-I SDSs (>1.96), numbers used for comparisons are variable (n = 16, 24, and 31). In the majority of cases, only one or two assays give different classification. There are more than three misclassified patients when comparing Liaison XL and iSYS, Liaison XL and Immulite, Liaison XL and Mediagnost ELISA, Liaison XL and Mediagnost RIA, and Mediagnost RIA and IGF-I RIACT. Finally, when assays give concordant results, they are more often in the high values of the techniques.

Discussion

Our results show significant variability among six commercial immunoassays for the determination of individual IGF-I SDS values and IGF-I classification of 102 patients with various GH disorders, despite the use of normative reference intervals obtained, for each of the six assays, in the same, large, well-selected population of healthy French adults (13), as recommended by the Consensus Group on the Standardization and Evaluation of GH and IGF-I Assays (6).

Reliable normative reference intervals are necessary for the diagnosis of acromegaly and GHD, for the follow-up of patients with GH disorders, and for the detection of remission and recurrence of GH-related diseases. In 2011, a consensus statement on the standardization and evaluation of GH and IGF-I assays proposed the use of the international recombinant IGF-I calibration standard preparation 02/254 and emphasized the importance of

Downloaded from https://academic.oup.com/jcem/article/102/8/2844/3819477 by guest on 01 June 2025

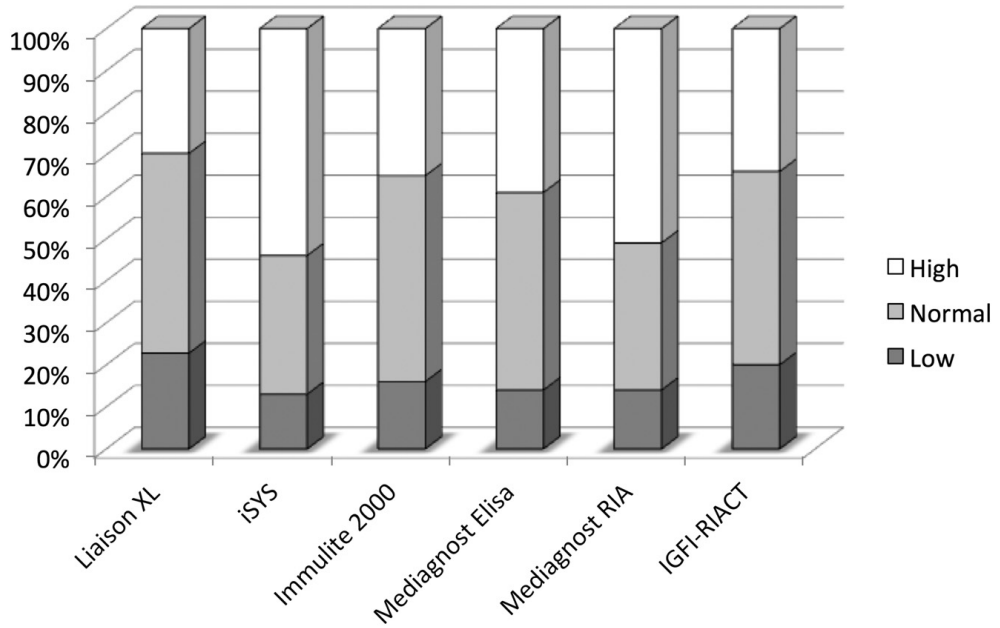

Figure 3. Percentages of patients with normal, low, and high IGF-I levels according to each of the six IGF-I immunoassays.

antibody specificity, quality control analysis, and the elimination of interference with binding proteins. It also emphasized the importance of obtaining normative data based on a random selection of individuals from the background population, representing all age groups, after exclusion of individuals with poorly controlled diabetes or renal or hepatic failure or taking medications (such as estrogen) that could affect outcome.

Based on this consensus statement, Bidlingmaier *et al.* (18) published normative data for the iSYS IGF-I assay obtained in a cohort of 15,014 healthy subjects, while we recently proposed IGF-I reference intervals obtained with six widely used immunoassays in the same population of 911 healthy French adults aged from 18 to 92 years, as per the consensus recommendations (13). The inclusion criteria were strict, with careful clinical evaluation, a medical history-taking that included ongoing treatments, and exclusion of subjects receiving steroid hormones. In addition, separate curves were constructed for each sex, in view of significantly different IGF-I levels between men and women. Normative data ranged between percentiles 2.5 and 97.5 and were reported in mass units and SDSs. Nevertheless, although we ensured the same preanalytical conditions for all six immunoassays, and although four of the six assays were calibrated against the same international reference standard 02/254, concordance across the assays remained variable, both for raw data and IGF-I SDSs (13).

To extend the results of our study of healthy individuals to the clinical setting, we created a group of patients of both sexes (57 males and 45 females) encompassing the whole spectrum of serum IGF-I levels, from very low (severe GHD) to very high (highly active acromegaly), representing the everyday practice of laboratories involved in IGF-I measurement. We therefore analyzed the concordance between the results obtained with each of the six assays in each of the 102 patients.

Pairwise agreement between the assays ranged from moderate to excellent. The best concordance was observed between iSYS and Mediagnost RIA. These two immunoassays, calibrated against the same international standard 02/254, classified fewer patients than the other four assays as "normal," and more patients as having "high" IGF-I serum levels. In the VARIETE study, the largest intercentile intervals and highest absolute values (in micrograms per liter) were obtained with Immulite and IGF-I RIACT, the two immunoassays calibrated against the old standard International Reference Preparation 87/518 (13). However, when using SDSs in the present group of patients, instead of absolute mass values, these two immunoassays classified similar percentages of patients as having "normal" IGF-I levels as the Liaison XL assay and Mediagnost ELISA. Moreover, the three automated assays (Immulite, Liaison XL, and iSYS) did not show excellent pairwise concordance: The pairs Liaison XL/iSYS and Immulite/iSYS

Downloaded from https://academic.oup.com/jcem/article/102/8/2844/3819477 by guest on 01 June 2025

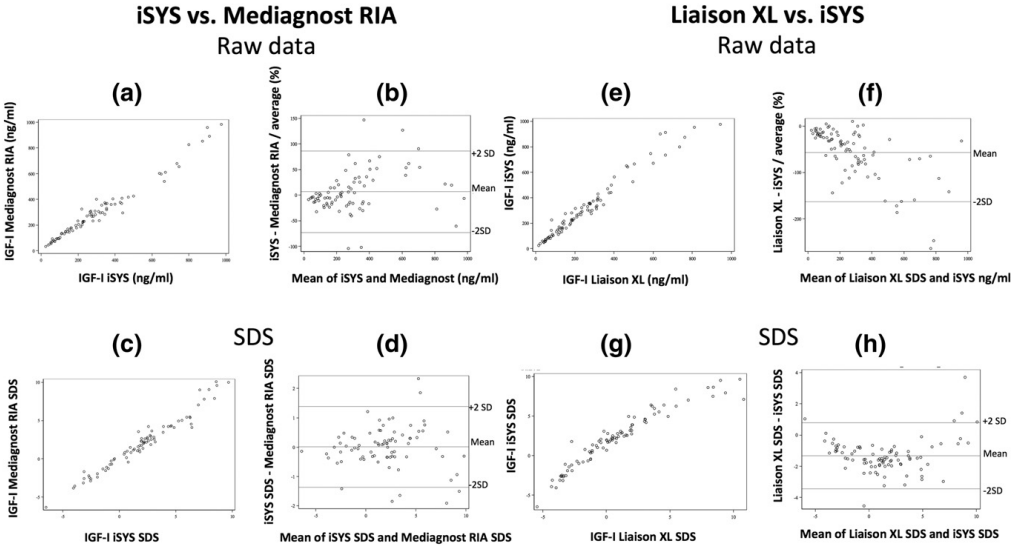

Figure 4. Comparisons between iSYS and Mediagnost RIA expressed as (a) scatter plots and (b) Bland-Altman plots for raw data, or (c) scatter plots and (d) Bland-Altman plots for SDSs, showing excellent overall agreement between the two immunoassays. Comparisons between Liaison XL and iSYS expressed as (e) scatter plots and (f) Bland-Altman plots for raw data, or (g) scatter plots and (h) Bland-Altman plots for SDSs, showing moderate overall agreement between these two immunoassays.

exhibited only moderate agreement (κ coefficients, 0.50 and 0.55, respectively), and only the pair Immulite/Liaison XL showed substantial agreement (κ coefficient, 0.76).

This lack of concordance between certain assays has already been reported (7–10): One possible explanation was that the populations used to establish normative data were different or that the quality of these normative data were suboptimal (too few patients studied, particularly in certain age ranges; bias; failure to select healthy subjects with regard to concurrent treatments or medical conditions interfering with IGF-I measurement; etc.) (5, 6, 11, 12, 19). This is why we used the same large healthy population to establish normative data for the six immunoassays used here. Despite this, discordant results persisted between some assays, with some pairs being clearly more discordant than others.

Another possible explanation for the lack of concordance is a difference in the technical procedure (5, 6, 12, 19). In the current study, the preanalytic procedure was exactly the same, and only the analytic procedure therefore differed. As underlined in our study of healthy volunteers, in which we also found such discordances (13), the most plausible explanation lies in the capacity of the assay to remove insulinlike growth factor binding proteins and the specificity and performance of the antibody. This may be particularly true for high IGF-I values, which are usually associated with high levels of insulinlike growth factor binding protein 3.

These results confirm that, even when using normative values established in the same population of healthy subjects, IGF-I results obtained with different assays in a given individual, whether healthy (as in the VARIETE study) or having a GH-related disorder (as in the current study), are sometimes very different, potentially leading to patient misclassification.

It is crucial to understand the reasons behind differences in the results of commercial IGF-I immunoassays. Assays with similar characteristics must be used for the follow-up of a given patient. Assays that tend to overestimate or

Table 2. Agreement Between IGF-I Assay Methods, Expressed as Weighted κ Coefficient

κ Coefficient	Liaison XL	iSYS	Immulite 2000	Mediagnost ELISA	Mediagnost RIA	IGF-I RIACT
Liaison XL		0.50	0.76	0.67	0.51	0.64
iSYS	0.50		0.55	0.69	0.81	0.71
Immulite 2000	0.76	0.55		0.77	0.62	0.65
Mediagnost ELISA	0.67	0.69	0.77		0.76	0.77
Mediagnost RIA	0.51	0.81	0.62	0.76		0.71
IGF-I RIACT	0.64	0.71	0.65	0.77	0.71	

Downloaded from https://academic.oup.com/jcem/article/102/8/2844/3819477 by guest on 01 June 2025

underestimate IGF-I values by comparison with other techniques must be clearly identified. Tandem liquid chromatography and mass spectrometry (LC-MS) may or may not prove to be a valid alternative (20, 21). Reference intervals obtained with LC-MS seem very similar to those obtained with immunoassays (22). However, LC-MS is a time consuming and complex method that requires expensive machines and technical expertise to control the many variables that can influence the results (23). Thus, despite their limitations, immunoassays will continue to be widely used, at least in the near future.

In conclusion, IGF-I levels obtained with six commercial IGF-I immunoassays widely used in clinical practice, and calculated IGF-I SDSs, were quite variable in patients with GH-related disorders, despite the use of normative reference intervals obtained in the same large, well-defined population of French healthy adults. It is not possible, according to the results of this study, to recommend one assay or the other. From a practical point of view, very high levels or very low levels of IGF-I are generally concordant, whatever the assay that is used, and classification of patients as having active acromegaly or severe GH deficiency is generally similar. On the contrary, when IGF-I levels are borderline, classification may differ from one assay to the other. This requires caution in interpretation of borderline IGF-I levels. In this context, we do not recommend to follow a patient and to take therapeutic decisions based on IGF-I SDSs calculated with one assay one day and another assay another day. On the contrary, a given patient should preferably be monitored with the same IGF-I assay.

Acknowledgments

We thank Céline Piedvache for her help in complementary statistical analysis.

Address all correspondence and requests for reprints to: Philippe Chanson, MD, Service d'Endocrinologie et des Maladies de la Reproduction, Hôpital de Bicêtre, 78 rue du Général Leclerc, 94275 Le Kremlin-Bicêtre, France. E-mail: philippe.chanson@bct.aphp.fr.

This work was supported by Programme Hospitalier de Recherche Clinique, French Ministry of Health Grant P081216/IDRCB 2009-A00892-55. M.M. received a grant from Fond National Suisse (Grant P1GEP3_155694) for her contribution to this work.

Clinical trial registry: ClinicalTrials.gov no. NCT01831648 (registered 3 April 2013).

Disclosure Summary: The authors have nothing to disclose.

References

1. Clemmons DR. Value of insulin-like growth factor system markers in the assessment of growth hormone status. *Endocrinol Metab Clin North Am.* 2007;**36**(1):109–129.
2. Katznelson L, Laws ER Jr, Melmed S, Molitch ME, Murad MH, Utz A, Wass JA; Endocrine Society. Acromegaly: an endocrine society clinical practice guideline. *J Clin Endocrinol Metab.* 2014; **99**(**11**):3933–3951.
3. Giustina A, Chanson P, Kleinberg D, Bronstein MD, Clemmons DR, Klibanski A, van der Lely AJ, Strasburger CJ, Lamberts SW, Ho KK, Casanueva FF, Melmed S; Acromegaly Consensus Group. Expert consensus document: a consensus on the medical treatment of acromegaly. *Nat Rev Endocrinol.* 2014;**10**(4): 243–248.
4. Molitch ME, Clemmons DR, Malozowski S, Merriam GR, Vance ML, Endocrine S; Endocrine Society. Evaluation and treatment of adult growth hormone deficiency: an Endocrine Society clinical practice guideline. *J Clin Endocrinol Metab.* 2011;**96**(6): 1587–1609.
5. Frystyk J, Freda P, Clemmons DR. The current status of IGF-I assays—a 2009 update. *Growth Horm IGF Res.* 2010;**20**(**1**): 8–18.
6. Clemmons DR. Consensus statement on the standardization and evaluation of growth hormone and insulin-like growth factor assays. *Clin Chem.* 2011;**57**(4):555–559.
7. Massart C, Poirier JY. Serum insulin-like growth factor-I measurement in the follow-up of treated acromegaly: comparison of four immunoassays. *Clin Chim Acta.* 2006;**373**(**1-2**): 176–179.
8. Pokrajac A, Wark G, Ellis AR, Wear J, Wieringa GE, Trainer PJ. Variation in GH and IGF-I assays limits the applicability of international consensus criteria to local practice. *Clin Endocrinol (Oxf).* 2007;**67**(**1**):65–70.
9. Krebs A, Wallaschofski H, Spilcke-Liss E, Kohlmann T, Brabant G, Völzke H, Nauck M. Five commercially available insulin-like growth factor I (IGF-I) assays in comparison to the former Nichols Advantage IGF-I in a growth hormone treated population. *Clin Chem Lab Med.* 2008;**46**(**12**):1776–1783.
10. Varewijck AJ, Lamberts SW, van der Lely AJ, Neggers SJ, Hofland LJ, Janssen JA. The introduction of the IDS-iSYS total IGF-**1** assay may have far-reaching consequences for diagnosis and treatment of GH deficiency. *J Clin Endocrinol Metab.* 2015;**100**(**1**): 309–316.
11. Brabant G, Wallaschofski H. Normal levels of serum IGF-I: determinants and validity of current reference ranges. *Pituitary.* 2007; **10**(2):129–133.
12. Junnila RK, Strasburger CJ, Bidlingmaier M. Pitfalls of insulin-like growth factor-I and growth hormone assays. *Endocrinol Metab Clin North Am.* 2015;**44**(**1**):27–34.
13. Chanson P, Arnoux A, Mavromati M, Brailly-Tabard S, Massart C, Young J, Piketty ML, Souberbielle JC, Investigators V; VARIETE Investigators. Reference values for IGF-I serum concentrations: comparison of six immunoassays. *J Clin Endocrinol Metab.* 2016; **101**(9):3450–3458.
14. Melmed S. Acromegaly. In: Melmed S, ed. The Pituitary. Malden, MA: Blackwell Science Inc.; 2011:463–474.
15. Chanson P, Salenave S, Kamenicky P. Acromegaly. *Handb Clin Neurol.* 2014;**124**:197–219.
16. Fermanian J. Measuring agreement between 2 observers: a quantitative case. *Rev Epidemiol Sante Publique.* 1984;**32**(6):408–413.
17. Fermanian J. Measurement of agreement between 2 judges. Qualitative cases. *Rev Epidemiol Sante Publique.* 1984;**32**(2): 140–147.
18. Bidlingmaier M, Friedrich N, Emeny RT, Spranger J, Wolthers OD, Roswall J, Körner A, Obermayer-Pietsch B, Hübener C, Dahlgren J, Frystyk J, Pfeiffer AF, Doering A, Bielohuby M, Wallaschofski H, Arafat AM. Reference intervals for insulin-like growth factor-**1** (IGF-I) from birth to senescence: results from a multicenter study using a new automated chemiluminescence IGF-I immunoassay conforming to recent international recommendations. *J Clin Endocrinol Metab.* 2014;**99**(5):1712–1721.
19. Clemmons DR. IGF-I assays: current assay methodologies and their limitations. *Pituitary.* 2007;**10**(2):121–128.

Downloaded from https://academic.oup.com/jcem/article/102/8/2844/3819477 by guest on 01 June 2025

20. Kay R, Halsall DJ, Annamalai AK, Kandasamy N, Taylor K, Fenwick S, Webb A, Wark G, Pleasance S, Gurnell M. A novel mass spectrometry-based method for determining insulin-like growth factor **1**: assessment in a cohort of subjects with newly diagnosed acromegaly. *Clin Endocrinol (Oxf).* 2013;**78**(3):424–430.
21. Ketha H, Singh RJ. Clinical assays for quantitation of insulin-like-growth-factor-**1** (IGF**1**). *Methods.* 2015;**81**:93–98.
22. Bystrom C, Sheng S, Zhang K, Caulfield M, Clarke NJ, Reitz R. Clinical utility of insulin-like growth factor **1** and 2; determination by high resolution mass spectrometry. *PLoS One.* 2012;7(9):e43457.
23. Hoofnagle AN, Whiteaker JR, Carr SA, Kuhn E, Liu T, Massoni SA, Thomas SN, Townsend RR, Zimmerman LJ, Boja E, Chen J, Crimmins DL, Davies SR, Gao Y, Hiltke TR, Ketchum KA, Kinsinger CR, Mesri M, Meyer MR, Qian WJ, Schoenherr RM, Scott MG, Shi T, Whiteley GR, Wrobel JA, Wu C, Ackermann BL, Aebersold R, Barnidge DR, Bunk DM, Clarke N, Fishman JB, Grant RP, Kusebauch U, Kushnir MM, Lowenthal MS, Moritz RL, Neubert H, Patterson SD, Rockwood AL, Rogers J, Singh RJ, Van Eyk JE, Wong SH, Zhang S, Chan DW, Chen X, Ellis MJ, Liebler DC, Rodland KD, Rodriguez H, Smith RD, Zhang Z, Zhang H, Paulovich AG. Recommendations for the generation, quantification, storage, and handling of peptides used for mass spectrometry-based assays. *Clin Chem.* 2016; **62**(**1**):48–69.

Downloaded from https://academic.oup.com/jcem/article/102/8/2844/3819477 by guest on 01 June 2025

6. THYROID NODULES

a. Introduction

i. *Epidemiology and classification of thyroid nodules*

Thyroid nodules are frequent in the adult general population, and their prevalence is higher in women and increases with age [109,110]. The exact prevalence varies in different studies, according to ethnicity and iodine status but is mostly related to detection method (autopsy, imaging, palpation) [111]. Autopsy and ultrasound series report a prevalence of thyroid nodules as high as 60%, while palpation series find a prevalence of 3-6% [109,111-113]. A recent meta-analysis including 102 epidemiological studies, estimated a total prevalence of thyroid nodules of 24.83% (95% CI 21.44-28.55) in the adult general population including all detection methods [114].

Thyroid nodules are increasingly recognized as incidentalomas due to the expanding use of imaging, such as ultrasound (US), computed tomography (CT-scan), magnetic resonance imaging (MRI) and positron emission tomography computed tomography (PET-CT) [115,116]. Still, most thyroid nodules are of benign nature and remain asymptomatic, while 10-13% regress spontaneously [117-119].

Malignancy rates of incidentally discovered nodules are estimated at 5-13% [118,120,121]. Malignancy rate is however higher for hypermetabolic thyroid incidentalomas in 18F-fluorodeoxyglucose PET (FDG-PET), calculated at 34.8% in a systematic review including more than 50,000 individuals with thyroid hypermetabolism on FDG-PET [122].

Among malignant thyroid nodules, 90% are well-differentiated thyroid carcinomas (75% papillary and 15% follicular carcinomas) and have an excellent prognosis, while more aggressive subtypes (oncocytic, medullary, anaplastic) are rare [123].

The incidence of thyroid cancer is steadily increasing, and this is most probably related to the increased detection rates of thyroid incidentalomas and papillary thyroid microcarcinomas [124]. However, papillary thyroid microcarcinomas defined to the thyroid gland are indolent tumors associated with no excess mortality if they do not present with lymph node involvement and distant metastasis and with a 0-5% recurrence rate in meta-analysis [125,126].

Thus, the diagnostic process of thyroid nodules must consider and weight the expected benefits of medical procedures including surgery with the disadvantages of overdiagnosis, patient anxiety and treatment related morbidity.

ii. *Diagnostic procedure and management of patients with thyroid nodules*

Upon diagnosis of a thyroid nodule, decision making must consider the circumstances of discovery, medical history and clinical examination findings [127-129]. Thus, rapid growth of a cervical mass, personal history of neck irradiation or whole-body irradiation for bone marrow transplant, especially during childhood, and family history of thyroid cancer or multiple endocrine neoplasia (MEN) syndromes should lead towards faster screening aiming to exclude malignancy. In addition, patient who present with a hard non-mobile neck mass, compression symptoms, enlarged cervical lymph nodes or vocal cord paralysis must be managed promptly due to increased probability of malignant thyroid neoplasm [129,130].

The next step after clinical evaluation and history taking is the determination of the thyroid nodule functional status by means of TSH measurement [127,129]. If TSH is low, which supports the diagnosis of hyperthyroidism, thyroid scintigraphy is indicated to determine if the nodule is hot (autonomous or hyperfunctioning) or cold. Autonomous thyroid nodules (hot nodules) are almost always benign while cold nodules necessitate ultrasound (US) evaluation for risk stratification [127,129].

Several US risk stratification systems (RSS) are available, such as the American College of Radiology Thyroid Imaging Reporting and Data System (ACR-TIRADS), the European TIRADS (EU-TIRADS), the American Thyroid Association RSS and the Korean TIRADS (K-TIRADS), while an International TIRADS is being prepared [127,131-135]. Their objective is to offer a systematic evaluation of suspicious ultrasound features and thus stratify the malignancy risk of thyroid nodules. The clinician can thus select those nodules that need to have fine needle aspiration (FNA) cytology.

The EU-TIRADS RSS is the most widely used in Europe in its 2017 version, while a 2023 update is available [127,132]. According to EU-TIRADS, the presence of any high-risk features on ultrasound (marked hypo-echogenicity, irregular margins, microcalcifications, non-oval shape) classifies a nodule in the highest risk category (EU-TIRADS 5). If none of those high-risk features is present, a nodule is classified as EU-TIRADS 4 if moderately hypo-echoic, as EU-TRADS 3 if iso-echoic or hyper-echoic and as EU-TIRADS 2 if anechoic or entirely spongiform. Thyroid FNA is indicated if an EU-TIRADS 5 nodule is larger than 10 mm in diameter. Size cut-offs for FNA for EU-TIRADS 4 and 3 nodules are > 15 mm and > 20 mm respectively, while FNA is not indicated for EU-TIRADS 2 nodules [127] (Table 1).

FNA cytology specimens are categorized according to the Bethesda System for reporting thyroid cytology in 6 categories: Bethesda I (non-diagnostic, with a ROM of 5-20%), Bethesda II (benign, with a ROM of 2-7%), Bethesda III (atypia of undetermined significance, with a ROM of 13-30%), Bethesda IV (follicular neoplasm, with a ROM of 23-

34%), Bethesda V (suspicious for malignancy, with a ROM 67-83%), and Bethesda VI (malignant, with a ROM 97-100%) [127,136]. While surveillance is suggested for nodules that are found to be benign on FNA cytology (Bethesda II), surgery is indicated for Bethesda V and VI nodules. For categories Bethesda III (confirmed twice) and IV, FNA cytology is considered non-diagnostic or indeterminate and management is more complex, with diagnostic lobectomy being the recommended option [127,129]. Still, due to the low ROM of these categories, most of Bethesda III or IV thyroid nodules, are found to be histologically benign after surgery.

US EU-TIRADS Category	US Characteristics	Size cut-offs for FNA	ROM
EU-TIRADS 2	anechoic or entirely spongiform	No FNA	≈0%
EU-TIRADS 3	isoechoic or hyperechoic	> 20 mm	2-4%
EU-TIRADS 4	moderately hypoechoic	> 15 mm	6-7%
EU-TIRADS 5	at least 1 high-risk feature among: marked hypo-echogenicity, irregular margins, microcalcifications, non-oval shape	> 10 mm	26-87%

Table 1: Ultrasound classification of thyroid nodules according to the EU-TIRADS system with size cut-odds for FNA and rates of malignancy (ROM)

iii. *Limitations in risk stratification of thyroid nodules and unnecessary procedures*

US RSS exhibit high sensitivity but low specificity. In a 2020 meta-analysis of 12 studies including 18 750 nodules that had FNA with the diagnosis of malignancy being confirmed by surgery, found sensitivities of 87% for the ATA RSS, 86% for the K-TIRADS, 74% for the ACR-TIRADS and 54% for the EU-TIRADS, while specificities were 31%, 28%, 64% and 53% respectively [137]. In a prospective study including 477 cases and evaluating the performance of different RSSs according to the rate of avoided unnecessary FNAs, the ACR-TIRADS RSS performed the best, classifying 53.4% of FNAs as unnecessary with a false negative rate as low as 2.2% [138].

Rate of malignancy (ROM) in the EU-TIRADS RSS is close de 0 for EU-TIRADS 2 nodules, 2-4% for EU-TIRADS 3, 6-17% for EU-TIRADS 4 and 26-87% for EU-TIRADS 5

nodules (Table 1) [132]. The wide ROM range in the EU-TIRADS 5 category (EU-TIRADS 2017) is related to low inter- and intra-observer reproducibility in assessing high-risk US features, and the definition of malignancy, either pathology, or cytology, in studies [139]. Reproducibility of RSSs also seems to be higher for high- and intermediate- suspicious nodules and worse for low-suspicious nodules [140]. The number of high-risk signs present for a nodule to be classified as EU-TIRADS 5 also helps to stratify ROM [140]. Thus, it would be helpful to have more reproductible RSSs for thyroid nodules with improved specificity in the high-suspicion categories, to better select nodules for FNA and avoid unnecessary procedures.

Various artificial intelligence (AI) software for thyroid ultrasonography are currently developed. Their objective is to optimize characterization of high-risk ultrasound features (marked hypo-echogenicity, microcalcifications, irregular margins, non-oval shape) and thus improve objectivity and precision [141,142]. Most AI software use static US images while some use video clips. A 2022 meta-analysis including 25 studies showed a sensitivity of 88% and specificity of 81% of AI-assisted diagnostic techniques for diagnosis of malignant versus benign nodules. AI-assisted thyroid US tools show similar diagnostic performance with expert clinicians and seem to particularly benefit the non-expert ultrasonographist, still, real-life utility and cost-effectiveness have not yet been proved [143,144].

Thyroid nodules with indeterminate results on FNA cytology (Bethesda III and IV) represent 10-40% of FNA procedures in literature and carry a ROM of 13-30% and 23-34% respectively [136]. As previously cited, guidelines recommend diagnostic lobectomy for those nodules with non-diagnostic FNA cytology, leading to a considerable number of unnecessary surgeries. If cancer is proved on final histology after lobectomy in these patients, and depending on staging and histological characteristics, some patients will need second surgery for completion thyroidectomy that could also be avoided if initial diagnosis was more precise. In a real data cohort from Geneva University Hospital (publication 4) including 1010 nodules in 862 patients subjected to FNA, unnecessary surgery, defined as surgery for indeterminate cytology with final benign histology or two-steps total thyroidectomy, was performed in more than half of patients who had surgery for a Bethesda III or IV nodules on FNA cytology. In this cohort, there were 174 (17.2%) Bethesda III and 168 (16.6%) Bethesda IV thyroid nodules but only 36% and 74% of them had surgery [145]. Final histology yielded benign results in 81% and 76% of operated Bethesda III and IV nodules respectively. Still, not all those surgeries could be considered unnecessary since some patients would have had surgery in any case because of local symptoms due to the size of the nodule. After excluding patients with local compressive symptoms, the rate of unnecessary surgery due to final benign histology was 53.5% for

Bethesda III and 64.9% for Bethesda IV nodules and unnecessary surgery due to two-step total thyroidectomy was 1.9% and 3.5% respectively [145].

Several preoperative molecular tests, among which the most widely studied ThyroSeq, and Afirma, have also been developed with the aim to assist selection of nodules with non-diagnostic FNA cytology (Bethesda III and IV) for surgery [146-148]. Molecular tests exhibit high specificity and negative predictive value which renders them valuable to use to rule-out malignancy and avoid unnecessary surgery. In a randomized clinical trial including 372 nodules with indeterminate cytology and comparing an RNA test (Afirma genomic sequencing classifier) with a DNA-RNA test (ThyroSeq v3 multigene genomic classifier), both tests showed high specificity (80% and 85% respectively) and high negative predictive value (100% and 99% respectively) for the diagnosis of malignancy and allowed 49% of patients to avoid diagnostic surgery [147]. Molecular tests are considered cost-effective in the US but despite their benefit for the individual patient, they are not covered by health insurance in Europe. Their cost-effectiveness remains to be proved in real-life settings and ideally considering the fact that not all patients with non-diagnostic FNA cytology will undergo surgery as suggested by current guidelines and that overtreatment could be an issue if every patient with a positive molecular test will have surgery.

b. Publications

Unnecessary thyroid surgery rate for suspicious nodule in the absence of molecular testing

Diagnostic evaluation and management of patients with thyroid nodules has limitations related to the insufficient specificity of ultrasound risk stratification scores that lead to unnecessary fine-needle aspiration (FNA) procedures. After FNA, cytology results are indeterminate (Bethesda III and IV), in 10-40% of patients. Rate of malignancy (ROM) in these non-diagnostic categories is 13-30% for Bethesda III nodules and 23-34% for Bethesda IV nodules and current guidelines recommend diagnostic lobectomy for those patients, which leads to unnecessary surgery.

We conducted a real data retrospective study including 1010 nodules in 862 patients (640 female) with a mean age of 54.2 years who had FNA cytology with rapid on-site evaluation (ROSE) between January 2017 and December 2021 in the endocrinology and radiology division of Geneva University Hospital.

In this cohort, 17.2% of nodules (n=174) were classified Bethesda III on FNA cytology and 16.6% (n=168) were classified Bethesda IV. Indeterminate cytology results concerned thus 33.8% of nodules that had FNA. Despite current recommendations for diagnostic lobectomy, only 36% of patients with Bethesda III nodules and 74% of patients with Bethesda IV nodules had surgery. Among nodules subjected to surgery because of non-diagnostic FNA cytology, 81% of Bethesda III nodules and 76% of Bethesda IV nodules were benign in histology. However, not all those could be considered unnecessary since some patients would have had surgery due to the presence of local compressive symptoms. In addition, some patients with non-diagnostic cytology and histology proved malignancy had subsequent second surgery for completion total thyroidectomy. After excluding patients who also had surgery because of local compressive symptoms, unnecessary surgery defined as final benign histology or two-step total thyroidectomy was performed in 56% of patients with Bethesda III nodules (54% final benign histology, 2% two-step total thyroidectomy) and 69% of patients with Bethesda IV nodules (65% final benign histology and 4% two-step thyroidectomy).

In this real-life cohort of thyroid nodules subjected to FNA cytology and without the aid of molecular testing, unnecessary surgery was performed in more than half of patients with non-diagnostic cytology (Bethesda III and IV).

European Thyroid JOURNAL | M Mavromati | 12:6 | e230114

RESEARCH

Unnecessary thyroid surgery rate for suspicious nodule in the absence of molecular testing

Maria Mavromati[1], Essia Saiji[2], Marco Stefano Demarchi[3], Vincent Lenoir[4], Amanda Seipel[2], Paulina Kuczma[3], François R Jornayvaz[1], Minerva Becker[4], Eugenio Fernandez[5], Claudio De Vito[2], Frédéric Triponez[3] and Sophie Leboulleux[1]

[1]Department of Endocrinology, Diabetology, Nutrition and Therapeutic Education, Geneva University Hospitals, Rue Gabrielle Perret Gentil, Geneva University, Geneva, Switzerland
[2]Department of Pathology, Geneva University Hospitals, Rue Gabrielle Perret Gentil, Geneva, Switzerland
[3]Department of Endocrine Surgery, Geneva University Hospitals, Rue Gabrielle Perret Gentil, Geneva, Switzerland
[4]Department of Radiology, Geneva University Hospitals, Rue Gabrielle Perret Gentil, Geneva, Switzerland
[5]Department of Oncology, Geneva University Hospitals, Rue Gabrielle Perret Gentil, Geneva, Switzerland

Correspondence should be addressed to M Mavromati: maria.mavromati@hcuge.ch

Abstract

Background: Molecular tests for suspicious thyroid nodules decrease rates of unnecessary surgeries but are not widely used due to reimbursement issues. The aim of this study was to assess the rate of unnecessary surgery performed in real-life setting for Bethesda III, IV and V nodules in the absence of molecular testing.
Method: This is a single-center retrospective study of consecutive patients undergoing fine needle aspiration cytology (FNAC) with rapid on-site evaluation between January 2017 and December 2021. Unnecessary surgery was defined as surgery performed because of Bethesda III, IV, or V results in the absence of local compressive symptoms with final benign pathology and as second surgery for completion thyroidectomy.
Results: In the 862 patients (640 females, mean age: 54.2 years), 1010 nodules (median size: 24.4 mm) underwent 1189 FNAC. Nodules were EU-TIRADS 2, 3, 4, and 5 in 3%, 34%, 42%, and 22% of cases, respectively. FNAC was Bethesda I, II, III, IV, V, and VI in 8%, 48%, 17%, 17%, 3%, and 6%, respectively. Surgery was performed in 36% of Bethesda III nodules (benign on pathology: 81%), in 74% of Bethesda IV nodules (benign on pathology: 76%) and in 97% of Bethesda V nodules (benign on pathology: 21%). Surgery was considered unnecessary in 56%, 68%, and 21% of patients with Bethesda III, IV, and V nodules, respectively.
Conclusion: In this real data cohort surgery was unnecessary in more than half of patients with Bethesda III and IV nodules and in 21% of patients with Bethesda V nodules.

Keywords

- Bethesda III
- Bethesda IV
- unnecessary thyroid surgery
- rate of malignancy

Introduction

Thyroid nodules are common, being palpable in 5% of adults and present in more than 60% of adults on high-frequency ultrasound (US) of the neck (1, 2). Five to 10% of these nodules are malignant, and patients are then treated with surgery, active surveillance, or local treatment, depending on the size of the nodule (3, 4, 5, 6). Given the high rate of thyroid nodules, most of which are benign, to reduce fine needle aspiration cytology (FNAC)

https://etj.bioscientifica.com
https://doi.org/10.1530/ETJ-23-0114
© 2023 the author(s)
Published by Bioscientifica Ltd.

This work is licensed under a Creative Commons Attribution 4.0 International License.

and unnecessary surgery, ultrasound scores are used to determine which nodules should undergo FNAC and the Bethesda classification is used to assess management based on the risk of malignancy (7, 8). Suspicious nodules with Bethesda III, Bethesda IV, and Bethesda V cytology, which represent 25–40% of all thyroid nodules and carry a risk of malignancy of 10–30%, 25–40%, and 50–75%, respectively, are often considered for diagnostic surgery. Furthermore, because malignancy is not diagnosed preoperatively, a lobectomy is the surgical procedure performed in most cases and depending on the final pathology and risk of recurrence classification, a second surgery may be necessary to complete total thyroidectomy. Preoperative molecular analyses are now available for suspicious nodules (9, 10, 11). With a sensitivity for cancer diagnosis of 91-95% and a specificity of 82–90%, their use reduces the rate of unnecessary surgery by 50% (12). Their benefit to the patient is obvious. However, they are not widely used in Europe due to their price and reimbursement issues. They are considered cost-effective in the U.S. through studies based on simulation cost analysis with theoretical models including thyroid nodule management based on guidelines and performance of molecular testing derived from clinical studies (13, 14, 15). Calculations of cost-effectiveness based on the assumption that, in the absence of molecular testing, all indeterminate thyroid nodules would be treated with diagnostic surgery are inaccurate because this assumption is erroneous. Furthermore, in this setting, if every positive molecular test leads to surgery, overtreatment could be an issue. Indeed, in a retrospective study comparing the management and cost of care of consecutive patients seen before and after the introduction of ThyroSeq v2, the rate and overtreatment only slightly decreased from 19% in the absence of molecular testing to 17% with the use of molecular testing with an overall rate of malignancy that remained equal and an average cost per thyroid cancer that increased by 47% (16). Although results would most probably differ with the use of more recent and accurate molecular tests, cost analysis should also consider follow-up after first management decisions. In order to evaluate the impact of molecular testing, assessing management in a real-world setting would be helpful.

The objectives of this study were to characterize consecutive nodules undergoing FNAC and to evaluate, in a real-life setting, the rate of unnecessary surgery performed for Bethesda III, IV, and V nodules in the absence of molecular testing.

Materials and methods

Study design and participants

This is a single-center retrospective study including data from consecutive patients who underwent ultrasound guided FNAC between January 2017 and December 2021 in the endocrinology and radiology division of Geneva University Hospital. Nodules diagnosed as intrathyroid metastases of nonthyroid malignancy and nodules with missing data on EU-TIRADS score and size were excluded.

The objectives of this study were to assess the rate of malignancy among nodules that underwent FNAC and among nodules that underwent surgery according to their EU-TIRADS and Bethesda results and to evaluate the rate of unnecessary surgery performed in a real-life setting for Bethesda III, IV, and V nodules in the absence of molecular testing.

The study was approved by the Swiss Ethics Committee in compliance with the Declaration of Helsinki; a waiver of informed consent was granted as the study was determined to involve no risk to the subjects included by using existing medical file information.

FNA cytology

FNA were performed in the endocrine and radiology divisions under US guidance. Nodules were classified according to the EU-TIRADS score (8).

Rapid on-site evaluation was provided in all cases, to determine validity of samples. FNAC samples were collected to prepare four to six slides immediately fixed in methanol for conventional smears. Material remaining in the needle was rinsed and collected in CytoLyt for ThinPrep slide preparation (Hologic, INC.). All conventional and liquid-based cytological smears were stained with Papanicolaou stain. Cytological features were evaluated and reported according to the criteria defined by The Bethesda System for Reporting Thyroid Cytopathology, second edition (17).

Surgical procedures

Indications for surgery included local symptoms and/or suspicious Bethesda III, IV, V, or VI results. In case of unilateral Bethesda III, IV, V nodules and in the absence of local compressive symptoms or voluminous contralateral nodules, a lobectomy was performed. In case of Bethesda VI result, in nodules of 2 cm or less and in the absence

https://etj.bioscientifica.com
https://doi.org/10.1530/ETJ-23-0114
© 2023 the author(s)
Published by Bioscientifica Ltd.

This work is licensed under a Creative Commons Attribution 4.0 International License.

of abnormal neck lymph node on US, a lobectomy with a prophylactic central ipsilateral neck lymph node dissection was performed. In case of Bethesda VI result in nodules larger than 2 cm, a total thyroidectomy with central ipsilateral neck lymph node was performed. Completion total thyroidectomy was proposed after lobectomy in case of cancer larger than 2 cm or metastatic lymph nodes.

Histology

Surgical specimens were formalin fixed and paraffin embedded. Nodules were classified according to the WHO 2017 criteria (18). For encapsulated neoplasms, the capsule was entirely submitted for histological examination. Diagnosis of noninvasive follicular thyroid neoplasms with papillary-like nuclear features (NIFTP) was based on WHO 2017 criteria (18). The regrouping of NIFTP, thyroid tumors of uncertain malignant potential (TUMP), and trabecular hyalinizing tumors within the low-risk neoplasms was done according to the WHO 2022 classification (19).

Data review

In the case of a first Bethesda I or Bethesda III cytology with repeated FNAC, the nodule was classified according to the results of the last cytology, with the exception of a last Bethesda I result, in which case the nodule remained in its first category. If no second FNAB was performed, the nodule was classified according to the only cytology result available.

Only the pathology result of the thyroid nodule that underwent FNAB was considered: if a cancer was incidentally found on pathology in addition to the benign nodule biopsied, the specimen was considered benign.

Evaluation criteria

The rate of malignancy for each EU-TIRADS category and each Bethesda class was calculated for nodules that underwent surgery and for all nodules that underwent FNAC. Nodules examined with FNAC without surgery were considered benign in this latter analysis. For EU-TIRADS 5 nodules, the rate of malignancy was also calculated according to the number of suspicious signs: solid hypoechoic, microcalcifications, irregular borders, taller than wide.

Unnecessary surgery was defined as surgery performed in patients without local compressive symptoms for Bethesda III, IV, or V nodules that were proved to be benign at final histology. A second surgery for completion thyroidectomy due to malignancy at initial lobectomy was also defined as unnecessary.

Statistical analysis

A descriptive analysis was done, with baseline characteristics reported as mean ± S.D., median (interquartile range), or number (%), as appropriate. Statistical analyses were performed in SPSS version X and in Stata version 17.0 SE. The association between EU-TIRADS score and rate of malignancy in Bethesda III, IV, and V nodules was evaluated with a Fisher's exact test. *P*-values <0.05 were considered statistically significant.

Results

Patients and nodules

Of the 870 patients who underwent FNAC between January 2017 and December 2021, 862 met the inclusion criteria (Fig. 1). In these patients, 1189 FNACs were performed in 1010 nodules (640 females, mean age: 54.2 years; range: 12.9–92).

Nodule characteristics, including size, EU-TIRADS score, and Bethesda classification, are detailed in Table 1. EU-TIRADS scores 3 and 4 were the most common among nodules (33.5% and 41.9%, respectively), and suspicious findings on the FNAC specimen, i.e. Bethesda III (17.2%), IV (16.6%), and V (3.0%), represented 36.8% of the total specimens.

Of the 100 nodules with a first Bethesda I result, 29 underwent a second procedure classified as Bethesda I in 12 cases, Bethesda II in 7 cases, Bethesda III in 9 cases, and Bethesda IV in 1 case. Of the 228 nodules with a first Bethesda III result, 114 underwent a second procedure

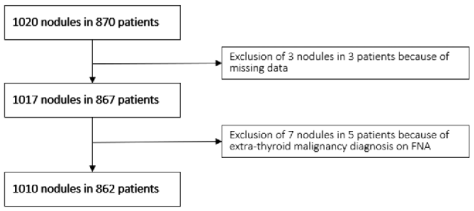

Figure 1
Flowchart of the nodules and patients included.

https://etj.bioscientifica.com
https://doi.org/10.1530/ETJ-23-0114
© 2023 the author(s)
Published by Bioscientifica Ltd.

This work is licensed under a Creative Commons Attribution 4.0 International License.

Table 1 Nodule characteristics. Data are presented as mean ± S.D. or as *n* (%).

Nodule characteristics	All	Nodules not operated	Nodules operated
Number of nodules	1010	669	341
Median size (mm)	24.4 ± 11.8	22 ± 11.9	21 ± 11.7
Laterality, *n* (%)			
Right	508 (50.3)	348 (52)	160 (46.9)
Left	465 (46)	299 (44.7)	166 (48.7)
Isthmus	37 (3.7)	22 (3.3)	15 (4.4)
EU-TIRADS score, *n* (%)			
2	30 (3.0)	25 (3.7)	5 (1.5)
3	338 (33.5)	253 (37.8)	85 (24.9)
4	423 (41.9)	277 (41.4)	146 (42.8)
5	219 (21.7)	114 (17.0)	105 (30.8)
Bethesda score, *n* (%)			
I	84 (8.3)	65 (9.7)	19 (5.6)
II	489 (48.4)	438 (65.5)	51 (15.0)
III	174 (17.2)	112 (16.7)	62 (18.2)
IV	168 (16.6)	44 (6.6)	124 (36.4)
V	30 (3.0)	1 (0.1)	29 (8.5)
VI	65 (6.4)	9 (1.3)	56 (16.4)

classified as Bethesda I in 7 cases, Bethesda II in 46 cases, Bethesda III in 45 cases, Bethesda IV in 13 cases, Bethesda V in 2 cases and Bethesda VI in 1 case.

Surgery

Surgery was performed in 287 (33.3%) patients with 341 (33.8%) nodules of which 36% were Bethesda III nodules, 74% were Bethesda IV nodules and 97% were Bethesda V nodules. Surgery consisted of lobectomy in 163 (56.8%) cases and total thyroidectomy in 124 (43.2%) cases. Based on the final pathology of the nodule, a completion thyroidectomy was performed in 12 cases. Final pathology was benign in 219 (64.2%) cases, low-risk neoplasm in 21 (6.2%) cases (NIFTP: 19 cases, TUMP: 1 case, hyalinizing trabecular tumor: 1 case) and malignant in 101 (29.6%) cases (papillary thyroid cancer: 79 cases; follicular thyroid cancer: 10 cases; oncocytic cancer: 5 cases; poorly differentiated thyroid cancer: 5 cases; medullary thyroid cancer: 1 case and intrathyroidal metastasis from renal carcinoma not diagnosed at cytology:1 case).

Rates of malignancy (malignancy only and malignancy plus low-risk neoplasm) in the operated and all nodules according to EU-TIRADS classification and Bethesda classification are detailed in Tables 2, 3, and 4 and Supplementary Table 1 (see section on supplementary materials given at the end of this article). Rate of malignancy in operated nodules, and including cancer only was 0% in EU-TIRADS 2 nodules, 11.9% in EU-TIRADS 3, 32.3% in EU-TIRADS 4, and 55.2% in EU-TIRADS 5 nodules. These rates were 0% for Bethesda I nodules, 2% for Bethesda II, 16.1% for Bethesda III, 14.5% for Bethesda IV, 58.6% for Bethesda V, and 98.2% for Bethesda VI nodules. Rates of malignancy, considering cancer only, among operated nodules that had 1, 2, 3, or 4 signs of EU-TIRADS 5 score were 48%, 44%, 75%, and 100%.

Combining EU-TIRADS score with Bethesda classification only slightly changed the rate of malignancy for Bethesda III, IV, and V nodules with a rate of malignancy in case of EUTIRADS 5 nodules compared to EUTIRADS 3 nodules of 7% vs 4% for Bethesda III cytology for all nodules undergoing FNAC and 19% vs 11% for all nodules undergoing surgery, 21% vs 14% for Bethesda IV cytology for all nodules undergoing FNAC and 29% vs. 21% for all nodules undergoing surgery, 75% vs 71% for Bethesda V cytology for all nodules undergoing FNAC

Table 2 Rates of malignancy (ROM) according to EU-TIRADS and Bethesda scores.

Nodule characteristics		Final pathology of nodules, *n*			ROM in operated nodules, %		ROM in FNAC nodules, %	
	n operated/total *n* (%)	Benign	LRN	Malignant	Malignant + LRN	Malignant only	Malignant + LRN	Malignant only
EU-TIRADS score								
2	5/30 (16.7)	5	0	0	0	0	0	0
3	85/338 (25.1)	70	5	10	17.6	11.9	4.4	3.0
4	146/423 (34.5)	102	11	33	30.1	32.3	10.4	7.8
5	105/219 (47.9)	42	5	58	60	55.2	28.8	26.5
Bethesda score								
I	19/84 (22.6)	19	0	0	0	0	0	0
II	51/489 (10.4)	50	0	1	2.0	2.0	0.2	0.2
III	62/174 (35.6)	50	2	10	19.3	16.1	6.9	5.7
IV	124/168 (73.8)	94	12	18	24.2	14.5	17.9	10.7
V	29/30 (96.7)	6	6	17	79.3	58.6	76.7	65.7
VI	56/65 (86.2)**	0	1*	55	100	98.2	86.2	84.6

*Trabecular hyalinizing tumor; **Reasons for the absence of surgery: anaplastic thyroid cancer on cytology (one case), active surveillance (one case), patient refusal (one case), and polymorbid condition or aggressive active extrathyroid malignancy (six patients).
LRN, low-risk neoplasm.

https://etj.bioscientifica.com
https://doi.org/10.1530/ETJ-23-0114
© 2023 the author(s)
Published by Bioscientifica Ltd.

This work is licensed under a Creative Commons Attribution 4.0 International License.

Table 3 Rates of malignancy (ROM) according to signs of EU-TIRADS 5 score.

Nodule characteristics		Final pathology of nodules, *n*			ROM in operated nodules		ROM in FNAB nodules	
	n operated/total *n* (%)	Benign	LRN*	Malignant	Malignant + LRN	Malignant only	Malignant + LRN	Malignant only
EU-TIRADS 5								
Taller than wide	16/37 (43.2%)	3	0	13	13/16 (81.3%)	13/16 (81.3%)	13/37 (35.1%)	13/37 (35.1%)
Irregular border	78/150 (52%)	30	2	46	48/78 (61.5%)	46/78 (59%)	48/150 (32%)	46/150 (30.7%)
Microcalcifications	36/57 (63.2%)	9	1	26	27/36 (75%)	26/36 (72.2%)	27/57 (47.4%)	26/57 (45.6%)
Hypoechoic	88/163 (54%)	35	5	48	53/88 (60.2%)	48/88 (54.5%)	53/163 (32.5%)	48/163 (29.4%)
EU-TIRADS 5								
1 sign	29/83 (35%)	13	2	14	16/29 (55.2%)	14/29 (48.3%)	16/83 (19.3%)	14/83 (16.9%)
2 signs	45/90 (50%)	22	3	20	23/45 (51.1%)	20/45 (44.4%)	23/90 (25.6%)	20/90 (22.2%)
3 signs	28/43 (65.1%)	7	0	21	21/28 (75%)	21/28 (75%)	21/43 (48.8%)	21/43 (48.8%)
4 signs	4/4 (100%)	0	0	4	4/4 (100%)	4/4 (100 %)	4/4 (100 %)	4/4 (100%)

*LRN: noninvasive follicular thyroid neoplasms with papillary-like features (NIFTP), thyroid tumors of uncertain malignant potential (TUMP), and trabecular hyalinizing tumors.
LRN, Low-risk neoplasm nodules.

and 75% vs 83% for all nodules undergoing surgery. Those changes were, however, nonsignificant when evaluated with Fisher's exact test (*P*-value: 0.7 for Bethesda III and IV classes, 1 for Bethesda V class.

Unnecessary surgery for Bethesda III, IV, and V nodules with final benign histology occurred in 28 (53.8%), 74 (64.9%), and four (16.7%) patients, respectively. Two-stage completion thyroidectomy after lobectomy for Bethesda III, IV, and V nodules was required in one (1.9%), four (3.5%), and one (4.2%) patients, respectively (Fig. 2).

Discussion

The malignancy rate of thyroid nodules varies from 1 to 10%, depending mainly on patient recruitment and the gold standard used to assess malignancy (20, 21, 22). In the absence of surgery, when the exclusion of malignancy is based on the absence of change in the thyroid nodule size on a 6-month follow-up neck US, the rate of malignancy is underestimated. When malignancy is considered in all Bethesda III, IV, V, or VI findings the rate of malignancy is, of course, overestimated. Finally, when malignancy rates are based on postoperative pathology, they are overestimated since nodules selected for surgery are more likely to be malignant and have a higher US rate for malignancy than nodules that do not undergo surgery. Knowing these difficulties, we chose, in our study, to calculate the rate of malignancy in all nodules undergoing FNAC and also only among nodules that underwent surgery. We also chose to calculate rates of malignancy taking into account cancer only, as well as cancer and low-risk neoplasms (including NIFTP and

Table 4 ROM (cancer + low risk neoplasm) according to EU-TIRADS score and Bethesda classification in operated nodules and FNAC nodules.

Bethesda score		EU-TIRADS score				
	n	All	2	3	4	5
Operated nodules						
I	19	0/19 (0%)	0/2 (0%)	0/5 (0%)	0/8 (0%)	0/4 (0%)
II	51	1/51 (2%)	0/2 (0%)	0/20 (0%)	1/26 (3.8%)	0/3 (0%)
III	62	12/62 (19.4%)	0/1 (0%)	2/19 (10.5%)	7/26 (26.9%)	3/16 (18.8%)
IV	124	30/124 (24.2%)	–	7/34 (20.6%)	15/62 (24.2%)	8/28 (28.6%)
V	29	23/29 (79.3%)	–	5/6 (83.3%)	12/15 (80%)	6/8 (75%)
VI	56	56/56 (100%)	–	1/1 (100%)	9/9 (100%)	46/46 (100%)
FNAC nodules						
I	84	0/84 (0%)	0/15 (0%)	0/23 (0%)	0/28 (0%)	0/18 (0%)
II	489	1/489 (0.2%)	0/14 (0%)	0/208 (0%)	1/209 (0.5%)	0/58 (0%)
III	174	12/174 (6.9%)	0/1 (0%)	2/50 (4%)	7/79 (8.9%)	3/44 (6.8%)
IV	168	30/168 (17.9%)	–	7/49 (14.3%)	15/81 (18.5%)	8/38 (21.1%)
V	30	23/30 (76.7%)	–	5/7 (71.4%)	12/15 (80%)	6/8 (75%)
VI	65	56/65 (86.2%)	–	1/1 (100%)	9/11 (81.8%)	46/53 (86.8%)

https://etj.bioscientifica.com
https://doi.org/10.1530/ETJ-23-0114
© 2023 the author(s)
Published by Bioscientifica Ltd.
This work is licensed under a Creative Commons Attribution 4.0 International License.

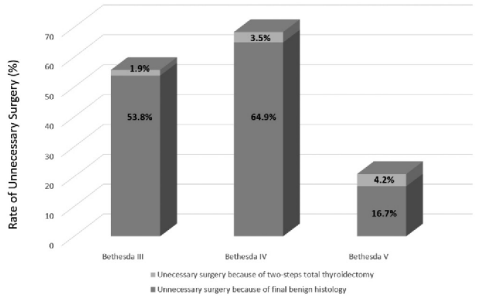

Figure 2
Rates of unnecessary surgery.

TUMP), since recommendations are to operate these tumors (19).

The malignancy rates in this study are comparable to those reported in the literature (17). In EU-TIRADS 2 nodules, the rate of malignancy (including cancer and low-risk neoplasms) was 0%. In EU-TIRADS 3 nodules, the rate of malignancy was 18% among nodules undergoing surgery and 4% among all nodules undergoing FNAC. These rates were 30% and 10% respectively, for EU-TIRADS 4 nodules, and 60% and 29%, respectively, for EU-TIRADS 5 nodules. In Bethesda I nodules, the rate of malignancy (including cancer and low-risk neoplasms) was 0%. In Bethesda II nodules, this rate was 2% among nodules undergoing surgery and 0.2% among all nodules undergoing FNAC. These rates were 19% and 7% respectively in Bethesda III nodules, 24% and 18%, respectively in Bethesda IV nodules, 79% and 77% respectively in Bethesda V nodules and finally 100% and 86%, respectively, in Bethesda VI nodules. As the rate of malignancy increased from Bethesda III to Bethesda V cytology, the percentage of nodules undergoing surgery increased, and the rates of malignancy among nodules undergoing surgery and among all nodules undergoing FNAC were closer. The situation was different for Bethesda VI nodules because some of the nodules were proposed for active surveillance.

The percentages of Bethesda I and Bethesda III nodules in our cohort were 8.3% and 17.2% respectively, similar to results from literature (5–11% and 2–18% respectively) (4). Rapid on-side evaluation of cytology specimens (ROSE) is found to improve adequacy of samples, and it is systematically performed in our practice (23, 24).

The impact of EU-TIRADS score on malignancy rates in each Bethesda class is a controversial subject (8). There is evidence that specific radiologic features, such as microcalcifications and irregular margins, can improve the diagnostic ability of cytology, but that seems not to be the case for EU-TIRADS classification per se (25). In the present study, combining EU-TIRADS score with Bethesda classification only slightly changed the rate of malignancy for Bethesda III, IV, and V nodules. Those changes were, however, nonsignificant which could be explained by the small numbers of patients in each category.

Bethesda V or VI FNAC in EU-TIRADS 3 nodules is a rare situation, occurring in only 0.8% of nodules. Bethesda II FNAC in EU-TIRADS 5 nodules was more common but still not frequent occurring in only 5.7% of the cohort. This is the basis for the recommendation of a second FNAC in Bethesda II – EU-TIRADS 5 nodules (6, 8). However, the limitation of the EU-TIRADS 5 category is its wide range of malignancy rate. Nodules are classified as EU-TIRADS 5 if at least one of the following signs is present: irregular margins, microcalcification, marked hypoechogenicity, or higher than wide shape. The number of EU-TIRADS 5 signs is, however, highly informative, even though interobserver reproducibility is questionable, with a rate of malignancy of 48% among EU-TIRADS 5 nodules undergoing surgery if only one sign is present, versus 75%, if three signs are present and a rate of 100% if four signs are present (26, 27).

The evaluation of molecular testing costs based on mathematical models and hypotheses must consider the actual management of suspicious FNAC and not only the recommended management, especially in diseases with excellent prognosis, which is the case of most thyroid cancers. In the present cohort, surgery was performed in 36% of Bethesda III nodules, 74% of Bethesda IV nodules and 97% of Bethesda V nodules in the absence of molecular testing, consistent with an increasing risk with increasing Bethesda classification. This also shows that surveillance of Bethesda III and IV nodules is already suggested to patients and is not limited to nodules of 10 mm or less. However, whether this management is relevant remains to be proven. Unnecessary surgery for Bethesda III, IV, and V nodules with final benign histology occurred in 106 (28.5%), a relatively small number. It was more frequent in Bethesda IV category, since more patients underwent surgery in this category. However, 42% of Bethesda III, IV, and V nodules did not undergo surgery when surgery is usually recommended. This brings up the question of

https://etj.bioscientifica.com
https://doi.org/10.1530/ETJ-23-0114
© 2023 the author(s)
Published by Bioscientifica Ltd.

This work is licensed under a Creative Commons Attribution 4.0 International License.

the interest in doing FNAC if surgery is not performed even in case of suspicious results. However, a Bethesda II result is reassuring, and knowing the risk of malignancy is the basis for a concrete discussion with patients to decide on management. Of note, two-stage completion thyroidectomy because of after lobectomy was not frequent, being necessary in only 1.9%, 3.5%, and 4.2% of Bethesda III, IV and V nodules, respectively.

Limitations of the present study include its retrospective design, and the absence of follow-up in nonoperated Bethesda III, IV, and V patients. Regarding this latest point, though, assessing the absence of malignancy based on follow-up is never guaranteed to be correct given the slow rate of progression of most thyroid cancers.

In conclusion, in this real data cohort, surgery was unnecessary in more than half of the patients with Bethesda III and IV nodules operated and in 21% of the patients with Bethesda V nodules operated.

Supplementary materials
This is linked to the online version of the article at https://doi.org/10.1530/ETJ-23-0114.

Declaration of interest
F Triponez has received consulting fees from Medtronic and Fluoptics, not related to the present study. S Leboulleux has received consulting fees from Lilly, Bayer, EISAI, not related to the present study. All other authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

Funding
This research did not receive any specific grant from any funding agency in the public, commercial, or not-for-profit sector.

Author contribution statement
MM and SL conceived the study, collected the data, analyzed the data, and wrote the paper. ES, CDV, AS, MD, PK, and FT analyzed the data and revised the manuscript. EF and FJ revised the manuscript.

Acknowledgements
The authors acknowledge the contributions of the Clinical Research Center, Geneva University Hospitals, and Faculty of Medicine, Geneva.

References

1 Mazzaferri EL. Management of a solitary thyroid nodule. *New England Journal of Medicine* 1993 **328** 553–559. (https://doi.org/10.1056/NEJM199302253280807)
2 Guth S, Theune U, Aberle J, Galach A & Bamberger CM. Very high prevalence of thyroid nodules detected by high frequency (13 MHz) ultrasound examination. *European Journal of Clinical Investigation* 2009 **39** 699–706. (https://doi.org/10.1111/j.1365-2362.2009.02162.x)
3 Leboulleux S, Tuttle RM, Pacini F & Schlumberger M. Papillary thyroid microcarcinoma: time to shift from surgery to active surveillance? *Lancet. Diabetes and Endocrinology* 2016 **4** 933–942. (https://doi.org/10.1016/S2213-8587(16)30180-2)
4 Haugen BR, Alexander EK, Bible KC, Doherty GM, Mandel SJ, Nikiforov YE, Pacini F, Randolph GW, Sawka AM, Schlumberger M, *et al.* 2015 American Thyroid Association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: the American Thyroid Association guidelines task force on thyroid nodules and differentiated thyroid cancer. *Thyroid* 2016 **26** 1–133. (https://doi.org/10.1089/thy.2015.0020)
5 Mauri G, Hegedus L, Bandula S, Cazzato RL, Czarniecka A, Dudeck O, Fugazzola L, Netea-Maier R, Russ G, Wallin G, *et al.* European Thyroid Association and Cardiovascular and Interventional Radiological Society of Europe 2021 Clinical practice guideline for the use of minimally invasive treatments in malignant thyroid lesions. *European Thyroid Journal* 2021 **10** 185–197. (https://doi.org/10.1159/000516469)
6 Leboulleux S, Lamartina L, Lecornet Sokol E, Menegaux F, Leenhardt L & Russ G. SFE-AFCE-SFMN 2022 Consensus on the management of thyroid nodules: follow-up: how and how long? *Annales d'Endocrinologie* 2022 **83** 407–414. (https://doi.org/10.1016/j.ando.2022.10.010)
7 Hoang JK, Asadollahi S, Durante C, Hegedus L, Papini E & Tessler FN. An international survey on utilization of five thyroid nodule risk stratification systems: a needs assessment with future implications. *Thyroid* 2022 **32** 675–681. (https://doi.org/10.1089/thy.2021.0558)
8 Russ G, Bonnema SJ, Erdogan MF, Durante C, Ngu R & Leenhardt L. European Thyroid Association guidelines for ultrasound malignancy risk stratification of thyroid nodules in adults: the EU-TIRADS. *European Thyroid Journal* 2017 **6** 225–237. (https://doi.org/10.1159/000478927)
9 Lupo MA, Walts AE, Sistrunk JW, Giordano TJ, Sadow PM, Massoll N, Campbell R, Jackson SA, Toney N, Narick CM, *et al.* Multiplatform molecular test performance in indeterminate thyroid nodules. *Diagnostic Cytopathology* 2020 **48** 1254–1264. (https://doi.org/10.1002/dc.24564)
10 Steward DL, Carty SE, Sippel RS, Yang SP, Sosa JA, Sipos JA, Figge JJ, Mandel S, Haugen BR, Burman KD, *et al.* Performance of a Multigene Genomic Classifier in thyroid nodules with indeterminate cytology: a prospective blinded multicenter study. *JAMA Oncol*. 2019 **5** 204–212. (https://doi.org/10.1001/jamaoncol.2018.4616)
11 Patel KN, Angell TE, Babiarz J, Barth NM, Blevins T, Duh QY, Ghossein RA, Harrell RM, Huang J, Kennedy GC, *et al.* Performance of a genomic sequencing classifier for the preoperative diagnosis of cytologically indeterminate thyroid nodules. *JAMA Surgery* 2018 **153** 817–824. (https://doi.org/10.1001/jamasurg.2018.1153)
12 Livhits MJ, Zhu CY, Kuo EJ, Nguyen DT, Kim J, Tseng CH, Leung AM, Rao J, Levin M, Douek ML, *et al.* Effectiveness of molecular testing techniques for diagnosis of indeterminate thyroid nodules: a randomized clinical trial. *JAMA Oncology* 2021 **7** 70–77. (https://doi.org/10.1001/jamaoncol.2020.5935)
13 Nicholson KJ, Roberts MS, McCoy KL, Carty SE & Yip L. Molecular testing versus diagnostic lobectomy in Bethesda III/IV thyroid nodules: a cost-effectiveness analysis. *Thyroid* 2019 **29** 1237–1243. (https://doi.org/10.1089/thy.2018.0779)
14 Dharampal N, Smith K, Harvey A, Paschke R, Rudmik L & Chandarana S. Cost-effectiveness analysis of molecular testing for cytologically indeterminate thyroid nodules. *Journal of Otolaryngology – Head and Neck Surgery* 2022 **51** 46. (https://doi.org/10.1186/s40463-022-00604-7)
15 Rivas AM, Nassar A, Zhang J, Casler JD, Chindris AM, Smallridge R & Bernet V. Thyroseq®V2.0 Molecular testing. Molecular testing: a cost-effective approach for the evaluation of indeterminate thyroid nodules. *Endocrine Practice* 2018 **24** 780–788. (https://doi.org/10.4158/ep-2018-0212)

© 2023 the author(s)
Published by Bioscientifica Ltd.

This work is licensed under a Creative Commons Attribution 4.0 International License.

16 Fazeli SR, Zehr B, Amraei R, Toraldo G, Guan H, Kindelberger D, Lee S & Cerda S. ThyroSeq v2 testing: impact on cytologic diagnosis, management, and cost of care in patients with thyroid nodule. *Thyroid* 2020 **30** 1528–1534. (https://doi.org/10.1089/thy.2019.0191)
17 Cibas ES & Ali SZ. The 2017 Bethesda system for reporting thyroid cytopathology. *Thyroid* 2017 **27** 1341–1346. (https://doi.org/10.1089/thy.2017.0500)
18 Haugen BR, Sawka AM, Alexander EK, Bible KC, Caturegli P, Doherty GM, Mandel SJ, Morris JC, Nassar A, Pacini F, *et al.* American Thyroid Association guidelines on the management of thyroid nodules and differentiated thyroid cancer task force review and recommendation on the proposed renaming of encapsulated follicular variant papillary thyroid carcinoma without invasion to noninvasive follicular thyroid neoplasm with papillary-like nuclear features. *Thyroid* 2017 **27** 481–483. (https://doi.org/10.1089/thy.2016.0628)
19 Baloch ZW, Asa SL, Barletta JA, Ghossein RA, Juhlin CC, Jung CK, LoVolsi VA, Papotti MG, Sobrinho-Simoes M, Tallini G, *et al.* Overview of the 2022 WHO classification of thyroid neoplasms. *Endocrine Pathology* 2022 **33** 27–63. (https://doi.org/10.1007/s12022-022-09707-3)
20 Grussendorf M, Ruschenburg I & Brabant G. Malignancy rates in thyroid nodules: a long-term cohort study of 17,592 patients. *European Thyroid Journal* 2022 **11**. (https://doi.org/10.1530/ETJ-22-0027)
21 Frates MC, Benson CB, Doubilet PM, Kunreuther E, Contreras M, Cibas ES, Orcutt J, Moore FDJr, Larsen PR, Marqusee E, *et al.* Prevalence and distribution of carcinoma in patients with solitary and multiple thyroid nodules on sonography. *Journal of Clinical Endocrinology and Metabolism* 2006 **91** 3411–3417. (https://doi.org/10.1210/jc.2006-0690)
22 Angell TE, Maurer R, Wang Z, Kim MI, Alexander CA, Barletta JA, Benson CB, Cibas ES, Cho NL, Doherty GM, *et al.* A cohort analysis of clinical and ultrasound variables predicting cancer risk in 20,001 consecutive thyroid nodules. *Journal of Clinical Endocrinology and Metabolism* 2019 **104** 5665–5672. (https://doi.org/10.1210/jc.2019-00664)
23 Witt BL & Schmidt RL. Rapid onsite evaluation improves the adequacy of fine-needle aspiration for thyroid lesions: a systematic review and meta-analysis. *Thyroid* 2013 **23** 428–435. (https://doi.org/10.1089/thy.2012.0211)
24 Schmidt RL, Witt BL, Lopez-Calderon LE & Layfield LJ. The influence of rapid onsite evaluation on the adequacy rate of fine-needle aspiration cytology: a systematic review and meta-analysis. *American Journal of Clinical Pathology* 2013 **139** 300–308. (https://doi.org/10.1309/AJCPEGZMJKC42VUP)
25 Seminati D, Capitoli G, Leni D, Fior D, Vacirca F, Di Bella C, Galimberti S, L'Imperio V & Pagni F. Use of diagnostic criteria from ACR and EU-TIRADS systems to improve the performance of cytology in thyroid nodule triage. *Cancers* 2021 **13**. (https://doi.org/10.3390/cancers13215439)
26 Persichetti A, Di Stasio E, Coccaro C, Graziano F, Bianchini A, Di Donna V, Corsello S, Valle D, Bizzarri G, Frasoldati A, *et al.* Inter- and intraobserver agreement in the assessment of thyroid nodule ultrasound features and classification systems: a blinded multicenter study. *Thyroid* 2020 **30** 237–242. (https://doi.org/10.1089/thy.2019.0360)
27 Russ G, Trimboli P & Buffet C. The New Era of TIRADSs to stratify the risk of malignancy of thyroid nodules: strengths, weaknesses and pitfalls. *Cancers* 2021 **13**. (https://doi.org/10.3390/cancers13174316)

Received 8 June 2023
Accepted 30 August 2023
Available online 1 September 2023
Version of Record published 11 October 2023

https://etj.bioscientifica.com
https://doi.org/10.1530/ETJ-23-0114
© 2023 the author(s)
Published by Bioscientifica Ltd.

This work is licensed under a Creative Commons Attribution 4.0 International License.

7. CONCLUSIONS AND PERSPECTIVES

Management of patients with endocrine tumors of the pituitary and the thyroid gland is challenging, also related to their increased recognition with imaging in current practice. There is uncertainty concerning relative indications for surgery while biologic and imaging markers are imperfect.

Pituitary adenomas

Pituitary neuroendocrine tumors (pitNETs), previously named pituitary adenomas, appear to be more frequent than believed to be in the past and are increasingly recognized as incidentalomas. They are benign tumors but increase morbidity if they are clinically functioning (prolactinomas, GH-secreting adenomas causing acromegaly, ACTH-secreting adenomas causing Cushing's disease, TSH-secreting adenomas), if they cause hypopituitarism by compressing the normal pituitary cells or stalk, or if they are large enough to compress adjacent structures causing visual impairment or headaches [6,13]. Clinically non-functioning pitNETS (NFPAs) are the second most frequent tumor in the general population, and most frequent among men and older adults [11,12].

While surgery is the first-line treatment for functioning pit-NETs, follow-up is generally suggested for patients with NFPAs who do not present visual impairment or imminent visual threat due to close proximity of the tumor with optic nerves and chiasm[13]. The fact that hypopituitarism could be considered as a relative indication for surgery in these patients has always been controversial and real data were lacking. Recent studies have shed light on different aspects that must be evaluated to answer this question, among others the natural history of non-functioning macroadenomas, complications of transsphenoidal surgery, morbidity and mortality of hypopituitarism (tumor-related or postoperative) and the risk of hypopituitarism after surgery.

Non-functioning pituitary macroadenomas have a higher growth potential than microadenomas, with a growth velocity of 1 mm / year, while those being close to optic chiasm will more frequently grow, and rate of new pituitary dysfunction has been estimated at 12% per year [62,66]. Serious complications of transsphenoidal surgery are rare in expert hands but still not negligible, impacting $\leq$ 5% of cases [67]. Furthermore, patients with NFPAs have increased mortality compared to the general population, unrelated the occurrence of surgery,

which is more pronounced in women and patients treated for central adrenal insufficiency with supra-physiologic doses of glucocorticoids [79].

Several studies have also demonstrated limited risk of new hormonal deficiencies after transsphenoidal surgery for non-functioning pit-NET, while several hormonal axes recover with surgery. However, while one-center cohorts from tertiary academic centers, among which the Geneva cohort show higher chances of recovery than risks of new pituitary deficiencies with surgery, data from the Swedish registry, the larger cohort reported so far, show somewhat higher rates of new hypopituitarism, which could be related to the inclusion of neurosurgeons without specific expertise on pituitary surgery [80-89]. Furthermore, data on rates of new deficiencies and recovery in an axis-specific manner are inconsistent. Finally, the most systematically found marker predicting recovery after transsphenoidal surgery was the presence of preoperative hyperprolactinemia, which possibly means that hypopituitarism in these cases is more related to reversible stalk compression rather than destruction of normal pituitary cells [88]. Future studies should be performed in larger series with better biologic definition of deficiencies based on systematic patient screening and dynamic testing whenever indicated, and with the objective to provide data on rates of new hormonal impairment and recovery by axis. Challenges also include the identification of markers predicting recovery or new hormonal deficiency also in an axis-specific manner.

Accurate IGF-1 measurements are fundamental for the diagnosis and follow-up of patients with growth hormone disorders (acromegaly, GH-deficiency). Unfortunately, considerable variability exists among different commercially available immunoassays, and even among different laboratories using the same assay [94,95].To address this problem, the WHO consensus statement for GH and IGF-1 assay standardization recommends the use of the same calibrator (02/254, a recombinant international IGF-1 standard preparation) for calibration of all IGF-1 immunoassays as well as obtaining age-related reference normative values from a healthy general population [97].

Following the example of an international multicenter study including 15 041 healthy individuals from Europe, Canada and the USA, with the objective to develop reference IGF-1 valuers for the iSYS kit, an automated immunoassay by Immunodiagnostic Systems, the VARIETE study, a French multicenter cohort provided normative data for 6 widely used commercial assays [101,102]. The VARIETE cohort included 911 healthy adults representing all age groups who had a detailed medical history and clinical examination and were enrolled after exclusion of factors potentially impacting IGF-1 levels. IGF-1 was measured in a cross-sectional manner and with each one of the 6 immunoassays (iSYS, Liaison XL, Immunite, IGF-1 RIACT, Mediagnost ELISA and Mediagnost RIA). Since the distribution of IGF-1 raw data is not Gaussian in the general population, gender- and age-specific curves were

established for each immunoassay after Cox-Box power transformation, to allow SD scores calculation. A calculator was also issued which is available online. Still, despite being obtained from the same healthy population, concordance among assays both on raw data and SD scores were only moderate to good, while concordance with the reference normal values proposed by each kit manufacturer were poor [102]. The VARIETE normative data were tested in a cohort of 102 patients with growth hormones disorders (acromegaly, GH deficiency, suspected GH-disorder) who had IGF-1 measurement with the 6 immunoassays. Despite the fact that normative reference values were obtained from the same healthy population, concordance in the classification of patients in low, normal or high IGF-1 levels remained variable, especially for those close to normal range [103].

Variability among IGF-1 immunoassays in the classification of patients with GH disorders has a considerable impact in clinical practice since therapeutic decisions are influenced. The use of the same international standard for assay standardization and the development of age- and sex-specific reference normative data from the same well selected healthy population do not seem to completely solve this problem. Thus, variability among immunoassays could be related to other analytical differences such as epitope specificities and methods for IGF-BP elimination.

As a result, clinicians should be aware of the particularities of the assay they are using, and each patient should ideally be followed with the same assay or at least with assays sharing similar characteristics, especially if IGF-1 levels are close to normal range. IGF-1 measurements by liquid chromatography tandem mass spectrometry (LC-MS) methods are not less erratic and on top of being time consuming and expensive, they propose normative data that resemble those of immunoassays. Finally, future cohorts for the development of IGF-1 should include different ethnic groups since significant differences have been found among different geographic populations [104].

Thyroid nodules

Thyroid nodules are the most frequent endocrine tumor in the adult general population, with a prevalence as high as 60% in autopsy and imaging series[111,112]. Their diagnosis is often made incidentally, due to the wide use of imaging in current practice. However, only 5-13% of incidentally discovered thyroid nodules are malignant (34.8% if they are found to be hypermetabolic in PET-CT scans), most of which have an excellent prognosis with adequate care [118,120,122].

Except for autonomous thyroid nodules (hot nodules) that are detected on thyroid scintigraphy in patients with low TSH and which are almost always benign, after clinical evaluation and history taking, an ultrasound (US) evaluation is indicated to select those

nodules with the right combination of size and suspicious features that need to be subjected to fine needle aspiration (FNA) cytology [127].

The most widely used US risk stratification system (RSS) in Europe is the European Thyroid Imaging reporting and Data System (EU-TIRADS) RSS, in its 2017 or newer 2023 version [127,132]. Other RSSs exist, among which the American College of Radiology TIRADS (ACR-TIRADS), the American Thyroid Association (ATA) RSS and the Korean RSS (K-RSS) [131,133,134]. An International TIRADS system (I-TIRADS) is also under development [135]. The aim of US RSSs is to classify thyroid nodules based on their ultrasound features, and to propose categorization, thus stratifying the risk of malignancy.

RSSs exhibit high sensitivity but low specificity [137]. Important inter- and intra-observer variability is also accounted in the evaluation of high-risk ultrasound features, which further lowers specificity [139]. The EU-TIRADS RSS recognizes 4 high-risk features (marked hypo-echogenicity, irregular margins, microcalcifications, non-oval shape) and the presence of at least one of those features classifies a nodule to the highest category of risk (EU-TIRADS 5) [127,132]. Rate of malignancy (ROM) is 26-87% for EU-TIRADS 5 nodules, 6-17% for EU-TIRADS 4 nodules and < 5% for not suspicious EU-TIRADS 3 and EU-TIRADS 2 nodules [132]. Thus, ROM has a considerably wide range in high-suspicion nodules which leads to high rates of unnecessary FNA procedures.

Furthermore, after FNA, cytology specimens are classified according to the Bethesda System in 6 categories: suspicious for malignancy or malignant nodules (Bethesda V and VI), benign nodules (Bethesda II), non-diagnostic cytology (Bethesda I) and indeterminate / non-diagnostic cytology (Bethesda III and IV) [136]. Bethesda III and Bethesda IV nodules (indeterminate / non-diagnostic cytology) account for 10-40% of FNAs and have a ROM of 13-30% and 23-34% respectively [136]. Guidelines recommend diagnostic lobectomy for these patients, leading to frequent unnecessary surgery, since most nodules will be found benign on final cytology. In addition, if cancer is proved on final histology after diagnostic lobectomy for indeterminate cytology, some patients will need second completion thyroidectomy (two-step thyroidectomy) which could also have been avoided with better preoperative diagnostic tools. In a real world cohort from Geneva University Hospital including 1010 nodules in 862 patients that had FNA between January 2017 and December 2021 in the endocrinology and radiology division, unnecessary surgery was performed in more than half of patients that had thyroidectomy because of non-diagnostic FNA cytology results, which accounts for 12.4% (n=107) of patients who were subjected to FNA.

Several artificial intelligence (AI) software for thyroid ultrasound evaluation are under development, aiming to improve nodule classification to categories of risk with more precision and less variability [141-143]. Their performance is comparable to that of clinicians with thyroid US

expertise in terms of detecting malignancy, still, their utility in real-world clinical prospective series remains to be proved.

Finally, molecular tests are also available for thyroid cytology specimens exhibiting high specificities and high negative predictive values, which make them valuable for the management of patients with non-diagnostic thyroid cytology (Bethesda III and Bethesda IV nodules) [146-148]. They are unfortunately not routinely used in Europe due to reimbursement issues, and their cost-effectiveness also remains to be proved by future studies.

Taken together, improvement of ultrasound risk stratification systems with higher specificity for malignant tumors and real-life studies on molecular tests for thyroid cytology will reduce unnecessary fine-needle aspiration procedures and unnecessary surgery.

8. REFERENCES

1. Di Ieva A, Rotondo F, Syro LV, Cusimano MD, Kovacs K. Aggressive pituitary adenomas--diagnosis and emerging treatments. Nat Rev Endocrinol 2014;10(7):423-35. DOI: 10.1038/nrendo.2014.64.
2. Melmed S. Pituitary-Tumor Endocrinopathies. N Engl J Med 2020;382(10):937-950. DOI: 10.1056/NEJMra1810772.
3. Gittleman H, Ostrom QT, Farah PD, et al. Descriptive epidemiology of pituitary tumors in the United States, 2004-2009. J Neurosurg 2014;121(3):527-35. DOI: 10.3171/2014.5.JNS131819.
4. Scangas GA, Laws ER, Jr. Pituitary incidentalomas. Pituitary 2014;17(5):486-91. DOI: 10.1007/s11102-013-0517-x.
5. Tritos NA, Miller KK. Diagnosis and Management of Pituitary Adenomas: A Review. JAMA 2023;329(16):1386-1398. DOI: 10.1001/jama.2023.5444.
6. Asa SL, Mete O, Perry A, Osamura RY. Overview of the 2022 WHO Classification of Pituitary Tumors. Endocr Pathol 2022;33(1):6-26. DOI: 10.1007/s12022-022-09703-7.
7. Yavropoulou MP, Tsoli M, Barkas K, Kaltsas G, Grossman A. The natural history and treatment of non-functioning pituitary adenomas (non-functioning PitNETs). Endocr Relat Cancer 2020;27(10):R375-R390. DOI: 10.1530/ERC-20-0136.
8. Daly AF, Beckers A. The Epidemiology of Pituitary Adenomas. Endocrinol Metab Clin North Am 2020;49(3):347-355. DOI: 10.1016/j.ecl.2020.04.002.
9. Buurman H, Saeger W. Subclinical adenomas in postmortem pituitaries: classification and correlations to clinical data. Eur J Endocrinol 2006;154(5):753-8. DOI: 10.1530/eje.1.02107.
10. Ezzat S, Asa SL, Couldwell WT, et al. The prevalence of pituitary adenomas: a systematic review. Cancer 2004;101(3):613-9. DOI: 10.1002/cncr.20412.
11. Fernandez A, Karavitaki N, Wass JA. Prevalence of pituitary adenomas: a community-based, cross-sectional study in Banbury (Oxfordshire, UK). Clin Endocrinol (Oxf) 2010;72(3):377-82. DOI: 10.1111/j.1365-2265.2009.03667.x.
12. Agustsson TT, Baldvinsdottir T, Jonasson JG, et al. The epidemiology of pituitary adenomas in Iceland, 1955-2012: a nationwide population-based study. Eur J Endocrinol 2015;173(5):655-64. DOI: 10.1530/EJE-15-0189.
13. Freda PU, Beckers AM, Katznelson L, et al. Pituitary incidentaloma: an endocrine society clinical practice guideline. J Clin Endocrinol Metab 2011;96(4):894-904. DOI: 10.1210/jc.2010-1048.
14. Melmed S. Acromegaly pathogenesis and treatment. J Clin Invest 2009;119(11):3189-202. DOI: 10.1172/JCI39375.
15. Daly AF, Rixhon M, Adam C, Dempegioti A, Tichomirowa MA, Beckers A. High prevalence of pituitary adenomas: a cross-sectional study in the province of Liege, Belgium. J Clin Endocrinol Metab 2006;91(12):4769-75. DOI: 10.1210/jc.2006-1668.
16. Schneider HJ, Sievers C, Saller B, Wittchen HU, Stalla GK. High prevalence of biochemical acromegaly in primary care patients with elevated IGF-1 levels. Clin Endocrinol (Oxf) 2008;69(3):432-5. DOI: 10.1111/j.1365-2265.2008.03221.x.

17. Rosario PW. Frequency of acromegaly in adults with diabetes or glucose intolerance and estimated prevalence in the general population. Pituitary 2011;14(3):217-21. DOI: 10.1007/s11102-010-0281-0.
18. Katznelson L, Laws ER, Jr., Melmed S, et al. Acromegaly: an endocrine society clinical practice guideline. J Clin Endocrinol Metab 2014;99(11):3933-51. DOI: 10.1210/jc.2014-2700.
19. Holdaway IM, Rajasoorya C. Epidemiology of acromegaly. Pituitary 1999;2(1):29-41. DOI: 10.1023/a:1009965803750.
20. Rajasoorya C, Holdaway IM, Wrightson P, Scott DJ, Ibbertson HK. Determinants of clinical outcome and survival in acromegaly. Clin Endocrinol (Oxf) 1994;41(1):95-102. DOI: 10.1111/j.1365-2265.1994.tb03789.x.
21. Melmed S. Acromegaly and cancer: not a problem? J Clin Endocrinol Metab 2001;86(7):2929-34. DOI: 10.1210/jcem.86.7.7635.
22. Holdaway IM, Rajasoorya RC, Gamble GD. Factors influencing mortality in acromegaly. J Clin Endocrinol Metab 2004;89(2):667-74. DOI: 10.1210/jc.2003-031199.
23. Sherlock M, Ayuk J, Tomlinson JW, et al. Mortality in patients with pituitary disease. Endocr Rev 2010;31(3):301-42. DOI: 10.1210/er.2009-0033.
24. Maione L, Brue T, Beckers A, et al. Changes in the management and comorbidities of acromegaly over three decades: the French Acromegaly Registry. Eur J Endocrinol 2017;176(5):645-655. DOI: 10.1530/EJE-16-1064.
25. Esposito D, Ragnarsson O, Granfeldt D, Marlow T, Johannsson G, Olsson DS. Decreasing mortality and changes in treatment patterns in patients with acromegaly from a nationwide study. Eur J Endocrinol 2018;178(5):459-469. DOI: 10.1530/EJE-18-0015.
26. Xiao Z, Xiao P, Wang Y, Fang C, Li Y. Risk of cancer in acromegaly patients: An updated meta-analysis and systematic review. PLoS One 2023;18(11):e0285335. DOI: 10.1371/journal.pone.0285335.
27. Petrossians P, Borges-Martins L, Espinoza C, et al. Gross total resection or debulking of pituitary adenomas improves hormonal control of acromegaly by somatostatin analogs. Eur J Endocrinol 2005;152(1):61-6. DOI: 10.1530/eje.1.01824.
28. Melmed S, Bronstein MD, Chanson P, et al. A Consensus Statement on acromegaly therapeutic outcomes. Nat Rev Endocrinol 2018;14(9):552-561. DOI: 10.1038/s41574-018-0058-5.
29. Fleseriu M. Advances in the pharmacotherapy of patients with acromegaly. Discov Med 2014;17(96):329-38. (https://www.ncbi.nlm.nih.gov/pubmed/24979253).
30. Melmed S, Cook D, Schopohl J, Goth MI, Lam KS, Marek J. Rapid and sustained reduction of serum growth hormone and insulin-like growth factor-1 in patients with acromegaly receiving lanreotide Autogel therapy: a randomized, placebo-controlled, multicenter study with a 52 week open extension. Pituitary 2010;13(1):18-28. DOI: 10.1007/s11102-009-0191-1.
31. Colao A, Auriemma RS, Pivonello R. The effects of somatostatin analogue therapy on pituitary tumor volume in patients with acromegaly. Pituitary 2016;19(2):210-21. DOI: 10.1007/s11102-015-0677-y.
32. Colao A, Auriemma RS, Pivonello R, Kasuki L, Gadelha MR. Interpreting biochemical control response rates with first-generation somatostatin analogues in acromegaly. Pituitary 2016;19(3):235-47. DOI: 10.1007/s11102-015-0684-z.

33. Carmichael JD, Bonert VS, Nuno M, Ly D, Melmed S. Acromegaly clinical trial methodology impact on reported biochemical efficacy rates of somatostatin receptor ligand treatments: a meta-analysis. J Clin Endocrinol Metab 2014;99(5):1825-33. DOI: 10.1210/jc.2013-3757.
34. Gadelha MR, Bronstein MD, Brue T, et al. Pasireotide versus continued treatment with octreotide or lanreotide in patients with inadequately controlled acromegaly (PAOLA): a randomised, phase 3 trial. Lancet Diabetes Endocrinol 2014;2(11):875-84. DOI: 10.1016/S2213-8587(14)70169-X.
35. Colao A, Bronstein MD, Freda P, et al. Pasireotide versus octreotide in acromegaly: a head-to-head superiority study. J Clin Endocrinol Metab 2014;99(3):791-9. DOI: 10.1210/jc.2013-2480.
36. Fleseriu M, Dreval A, Bondar I, et al. Maintenance of response to oral octreotide compared with injectable somatostatin receptor ligands in patients with acromegaly: a phase 3, multicentre, randomised controlled trial. Lancet Diabetes Endocrinol 2022;10(2):102-111. DOI: 10.1016/S2213-8587(21)00296-5.
37. van der Lely AJ, Biller BM, Brue T, et al. Long-term safety of pegvisomant in patients with acromegaly: comprehensive review of 1288 subjects in ACROSTUDY. J Clin Endocrinol Metab 2012;97(5):1589-97. DOI: 10.1210/jc.2011-2508.
38. Freda PU, Gordon MB, Kelepouris N, Jonsson P, Koltowska-Haggstrom M, van der Lely AJ. Long-term treatment with pegvisomant as monotherapy in patients with acromegaly: experience from ACROSTUDY. Endocr Pract 2015;21(3):264-74. DOI: 10.4158/EP14330.OR.
39. Qiao N, He M, Shen M, et al. Comparative Efficacy of Medical Treatment for Acromegaly: A Systematic Review and Network Meta-Analysis of Integrated Randomized Trials and Observational Studies. Endocr Pract 2020;26(4):454-462. DOI: 10.4158/EP-2019-0528.
40. Knappe UJ, Petroff D, Quinkler M, et al. Fractionated radiotherapy and radiosurgery in acromegaly: analysis of 352 patients from the German Acromegaly Registry. Eur J Endocrinol 2020;182(3):275-284. DOI: 10.1530/EJE-19-0784.
41. Lacroix A, Feelders RA, Stratakis CA, Nieman LK. Cushing's syndrome. Lancet 2015;386(9996):913-27. DOI: 10.1016/S0140-6736(14)61375-1.
42. Lindholm J, Juul S, Jorgensen JO, et al. Incidence and late prognosis of cushing's syndrome: a population-based study. J Clin Endocrinol Metab 2001;86(1):117-23. DOI: 10.1210/jcem.86.1.7093.
43. Ragnarsson O, Olsson DS, Chantzichristos D, et al. The incidence of Cushing's disease: a nationwide Swedish study. Pituitary 2019;22(2):179-186. DOI: 10.1007/s11102-019-00951-1.
44. Dekkers OM, Horvath-Puho E, Jorgensen JO, et al. Multisystem morbidity and mortality in Cushing's syndrome: a cohort study. J Clin Endocrinol Metab 2013;98(6):2277-84. DOI: 10.1210/jc.2012-3582.
45. Ragnarsson O, Olsson DS, Papakokkinou E, et al. Overall and Disease-Specific Mortality in Patients With Cushing Disease: A Swedish Nationwide Study. J Clin Endocrinol Metab 2019;104(6):2375-2384. DOI: 10.1210/jc.2018-02524.
46. Balomenaki M, Vassiliadi DA, Tsagarakis S. Cushing's disease: risk of recurrence following trans-sphenoidal surgery, timing and methods for evaluation. Pituitary 2022;25(5):718-721. DOI: 10.1007/s11102-022-01226-y.

47. Petersenn S, Beckers A, Ferone D, et al. Therapy of endocrine disease: outcomes in patients with Cushing's disease undergoing transsphenoidal surgery: systematic review assessing criteria used to define remission and recurrence. Eur J Endocrinol 2015;172(6):R227-39. DOI: 10.1530/EJE-14-0883.
48. Pivonello R, Fleseriu M, Newell-Price J, et al. Efficacy and safety of osilodrostat in patients with Cushing's disease (LINC 3): a multicentre phase III study with a double-blind, randomised withdrawal phase. Lancet Diabetes Endocrinol 2020;8(9):748-761. DOI: 10.1016/S2213-8587(20)30240-0.
49. Gadelha M, Bex M, Feelders RA, et al. Randomized Trial of Osilodrostat for the Treatment of Cushing Disease. J Clin Endocrinol Metab 2022;107(7):e2882-e2895. DOI: 10.1210/clinem/dgac178.
50. Mehta GU, Ding D, Patibandla MR, et al. Stereotactic Radiosurgery for Cushing Disease: Results of an International, Multicenter Study. J Clin Endocrinol Metab 2017;102(11):4284-4291. DOI: 10.1210/jc.2017-01385.
51. Auriemma RS, Pirchio R, Pivonello C, Garifalos F, Colao A, Pivonello R. Approach to the Patient With Prolactinoma. J Clin Endocrinol Metab 2023;108(9):2400-2423. DOI: 10.1210/clinem/dgad174.
52. Petersenn S, Fleseriu M, Casanueva FF, et al. Diagnosis and management of prolactin-secreting pituitary adenomas: a Pituitary Society international Consensus Statement. Nat Rev Endocrinol 2023;19(12):722-740. DOI: 10.1038/s41574-023-00886-5.
53. Colao A, di Sarno A, Pivonello R, di Somma C, Lombardi G. Dopamine receptor agonists for treating prolactinomas. Expert Opin Investig Drugs 2002;11(6):787-800. DOI: 10.1517/13543784.11.6.787.
54. Maiter D. Management of Dopamine Agonist-Resistant Prolactinoma. Neuroendocrinology 2019;109(1):42-50. DOI: 10.1159/000495775.
55. Souteiro P, Karavitaki N. Dopamine agonist resistant prolactinomas: any alternative medical treatment? Pituitary 2020;23(1):27-37. DOI: 10.1007/s11102-019-00987-3.
56. Souteiro P, Belo S, Carvalho D. Dopamine agonists in prolactinomas: when to withdraw? Pituitary 2020;23(1):38-44. DOI: 10.1007/s11102-019-00989-1.
57. Raverot G, Dantony E, Beauvy J, et al. Risk of Recurrence in Pituitary Neuroendocrine Tumors: A Prospective Study Using a Five-Tiered Classification. J Clin Endocrinol Metab 2017;102(9):3368-3374. DOI: 10.1210/jc.2017-00773.
58. Trouillas J, Jaffrain-Rea ML, Vasiljevic A, Raverot G, Roncaroli F, Villa C. How to Classify the Pituitary Neuroendocrine Tumors (PitNET)s in 2020. Cancers (Basel) 2020;12(2). DOI: 10.3390/cancers12020514.
59. Galm BP, Martinez-Salazar EL, Swearingen B, et al. MRI texture analysis as a predictor of tumor recurrence or progression in patients with clinically non-functioning pituitary adenomas. Eur J Endocrinol 2018;179(3):191-198. DOI: 10.1530/EJE-18-0291.
60. Hamblin R, Fountas A, Lithgow K, et al. Natural history of non-functioning pituitary microadenomas: results from the UK non-functioning pituitary adenoma consortium. Eur J Endocrinol 2023;189(1):87-95. DOI: 10.1093/ejendo/lvad070.
61. Dekkers OM, Pereira AM, Romijn JA. Treatment and follow-up of clinically nonfunctioning pituitary macroadenomas. J Clin Endocrinol Metab 2008;93(10):3717-26. DOI: 10.1210/jc.2008-0643.
62. Castinetti F, Dufour H, Gaillard S, et al. Non-functioning pituitary adenoma: when and how to operate? What pathologic criteria for typing? Ann Endocrinol (Paris) 2015;76(3):220-7. DOI: 10.1016/j.ando.2015.04.007.

63. Sam AH, Shah S, Saleh K, et al. Clinical outcomes in patients with nonfunctioning pituitary adenomas managed conservatively. Clin Endocrinol (Oxf) 2015;83(6):861-5. DOI: 10.1111/cen.12860.
64. Mete O, Lopes MB. Overview of the 2017 WHO Classification of Pituitary Tumors. Endocr Pathol 2017;28(3):228-243. DOI: 10.1007/s12022-017-9498-z.
65. Arita K, Tominaga A, Sugiyama K, et al. Natural course of incidentally found nonfunctioning pituitary adenoma, with special reference to pituitary apoplexy during follow-up examination. J Neurosurg 2006;104(6):884-91. DOI: 10.3171/jns.2006.104.6.884.
66. Fernandez-Balsells MM, Murad MH, Barwise A, et al. Natural history of nonfunctioning pituitary adenomas and incidentalomas: a systematic review and metaanalysis. J Clin Endocrinol Metab 2011;96(4):905-12. DOI: 10.1210/jc.2010-1054.
67. Murad MH, Fernandez-Balsells MM, Barwise A, et al. Outcomes of surgical treatment for nonfunctioning pituitary adenomas: a systematic review and meta-analysis. Clin Endocrinol (Oxf) 2010;73(6):777-91. DOI: 10.1111/j.1365-2265.2010.03875.x.
68. Gondim JA, Almeida JP, Albuquerque LA, et al. Endoscopic endonasal approach for pituitary adenoma: surgical complications in 301 patients. Pituitary 2011;14(2):174-83. DOI: 10.1007/s11102-010-0280-1.
69. Kim JH, Lee JH, Lee JH, Hong AR, Kim YJ, Kim YH. Endoscopic Transsphenoidal Surgery Outcomes in 331 Nonfunctioning Pituitary Adenoma Cases After a Single Surgeon Learning Curve. World Neurosurg 2018;109:e409-e416. DOI: 10.1016/j.wneu.2017.09.194.
70. Messerer M, De Battista JC, Raverot G, et al. Evidence of improved surgical outcome following endoscopy for nonfunctioning pituitary adenoma removal. Neurosurg Focus 2011;30(4):E11. DOI: 10.3171/2011.1.FOCUS10308.
71. Ammirati M, Wei L, Ciric I. Short-term outcome of endoscopic versus microscopic pituitary adenoma surgery: a systematic review and meta-analysis. J Neurol Neurosurg Psychiatry 2013;84(8):843-9. DOI: 10.1136/jnnp-2012-303194.
72. Messerer M, Dubourg J, Raverot G, et al. Non-functioning pituitary macro-incidentalomas benefit from early surgery before becoming symptomatic. Clin Neurol Neurosurg 2013;115(12):2514-20. DOI: 10.1016/j.clineuro.2013.10.007.
73. Losa M, Donofrio CA, Barzaghi R, Mortini P. Presentation and surgical results of incidentally discovered nonfunctioning pituitary adenomas: evidence for a better outcome independently of other patients' characteristics. Eur J Endocrinol 2013;169(6):735-42. DOI: 10.1530/EJE-13-0515.
74. Olsson DS, Nilsson AG, Bryngelsson IL, Trimpou P, Johannsson G, Andersson E. Excess Mortality in Women and Young Adults With Nonfunctioning Pituitary Adenoma: A Swedish Nationwide Study. J Clin Endocrinol Metab 2015;100(7):2651-8. DOI: 10.1210/jc.2015-1475.
75. Ntali G, Capatina C, Fazal-Sanderson V, et al. Mortality in patients with non-functioning pituitary adenoma is increased: systematic analysis of 546 cases with long follow-up. Eur J Endocrinol 2016;174(2):137-45. DOI: 10.1530/EJE-15-0967.
76. Zueger T, Kirchner P, Herren C, et al. Glucocorticoid replacement and mortality in patients with nonfunctioning pituitary adenoma. J Clin Endocrinol Metab 2012;97(10):E1938-42. DOI: 10.1210/jc.2012-2432.
77. Hammarstrand C, Ragnarsson O, Bengtsson O, Bryngelsson IL, Johannsson G, Olsson DS. Comorbidities in patients with non-functioning pituitary adenoma: influence of

long-term growth hormone replacement. Eur J Endocrinol 2018;179(4):229-237. DOI: 10.1530/EJE-18-0370.
78. O'Reilly MW, Reulen RC, Gupta S, et al. ACTH and gonadotropin deficiencies predict mortality in patients treated for nonfunctioning pituitary adenoma: long-term follow-up of 519 patients in two large European centres. Clin Endocrinol (Oxf) 2016;85(5):748-756. DOI: 10.1111/cen.13141.
79. Tampourlou M, Fountas A, Ntali G, Karavitaki N. Mortality in patients with non-functioning pituitary adenoma. Pituitary 2018;21(2):203-207. DOI: 10.1007/s11102-018-0863-9.
80. Webb SM, Rigla M, Wagner A, Oliver B, Bartumeus F. Recovery of hypopituitarism after neurosurgical treatment of pituitary adenomas. J Clin Endocrinol Metab 1999;84(10):3696-700. DOI: 10.1210/jcem.84.10.6019.
81. Nelson AT, Jr., Tucker HS, Jr., Becker DP. Residual anterior pituitary function following transsphenoidal resection of pituitary macroadenomas. J Neurosurg 1984;61(3):577-80. DOI: 10.3171/jns.1984.61.3.0577.
82. Nomikos P, Ladar C, Fahlbusch R, Buchfelder M. Impact of primary surgery on pituitary function in patients with non-functioning pituitary adenomas -- a study on 721 patients. Acta Neurochir (Wien) 2004;146(1):27-35. DOI: 10.1007/s00701-003-0174-3.
83. Harary M, DiRisio AC, Dawood HY, et al. Endocrine function and gland volume after endoscopic transsphenoidal surgery for nonfunctional pituitary macroadenomas. J Neurosurg 2019;131(4):1142-1151. DOI: 10.3171/2018.5.JNS181054.
84. Alexopoulou O, Everard V, Etoa M, et al. Outcome of pituitary hormone deficits after surgical treatment of nonfunctioning pituitary macroadenomas. Endocrine 2021;73(1):166-176. DOI: 10.1007/s12020-021-02701-5.
85. Hwang JY, Aum DJ, Chicoine MR, et al. Axis-specific analysis and predictors of endocrine recovery and deficits for non-functioning pituitary adenomas undergoing endoscopic transsphenoidal surgery. Pituitary 2020;23(4):389-399. DOI: 10.1007/s11102-020-01045-z.
86. Jahangiri A, Wagner JR, Han SW, et al. Improved versus worsened endocrine function after transsphenoidal surgery for nonfunctional pituitary adenomas: rate, time course, and radiological analysis. J Neurosurg 2016;124(3):589-95. DOI: 10.3171/2015.1.JNS141543.
87. Mavromati M, Mavrakanas T, Jornayvaz FR, et al. The impact of transsphenoidal surgery on pituitary function in patients with non-functioning macroadenomas. Endocrine 2023;81(2):340-348. DOI: 10.1007/s12020-023-03400-z.
88. Arafah BM, Prunty D, Ybarra J, Hlavin ML, Selman WR. The dominant role of increased intrasellar pressure in the pathogenesis of hypopituitarism, hyperprolactinemia, and headaches in patients with pituitary adenomas. J Clin Endocrinol Metab 2000;85(5):1789-93. DOI: 10.1210/jcem.85.5.6611.
89. Al-Shamkhi N, Berinder K, Borg H, et al. Pituitary function before and after surgery for nonfunctioning pituitary adenomas-data from the Swedish Pituitary Register. Eur J Endocrinol 2023;189(2):217-224. DOI: 10.1093/ejendo/lvad104.
90. Pedersen MB, Dukanovic S, Springborg JB, Andreassen M, Krogh J. Endocrine Function after Transsphenoidal Surgery in Patients with Non-Functioning Pituitary Adenomas: A Systematic Review and Meta-Analysis. Neuroendocrinology 2022;112(9):823-834. DOI: 10.1159/000522090.

91. Giustina A, Biermasz N, Casanueva FF, et al. Consensus on criteria for acromegaly diagnosis and remission. Pituitary 2024;27(1):7-22. DOI: 10.1007/s11102-023-01360-1.
92. Fleseriu M, Hashim IA, Karavitaki N, et al. Hormonal Replacement in Hypopituitarism in Adults: An Endocrine Society Clinical Practice Guideline. J Clin Endocrinol Metab 2016;101(11):3888-3921. DOI: 10.1210/jc.2016-2118.
93. Yuen KCJ, Biller BMK, Radovick S, et al. American Association of Clinical Endocrinologists and American College of Endocrinology Guidelines for Management of Growth Hormone Deficiency in Adults and Patients Transitioning from Pediatric to Adult Care. Endocr Pract 2019;25(11):1191-1232. DOI: 10.4158/GL-2019-0405.
94. Clemmons DR, Bidlingmaier M. IGF-I assay methods and biologic variability: evaluation of acromegaly treatment response. Eur J Endocrinol 2024;191(1):R1-R8. DOI: 10.1093/ejendo/lvae065.
95. Pokrajac A, Wark G, Ellis AR, Wear J, Wieringa GE, Trainer PJ. Variation in GH and IGF-I assays limits the applicability of international consensus criteria to local practice. Clin Endocrinol (Oxf) 2007;67(1):65-70. DOI: 10.1111/j.1365-2265.2007.02836.x.
96. Burns C, Rigsby P, Moore M, Rafferty B. The First International Standard For Insulin-like Growth Factor-1 (IGF-1) for immunoassay: preparation and calibration in an international collaborative study. Growth Horm IGF Res 2009;19(5):457-62. DOI: 10.1016/j.ghir.2009.02.002.
97. Clemmons DR. Consensus statement on the standardization and evaluation of growth hormone and insulin-like growth factor assays. Clin Chem 2011;57(4):555-9. DOI: 10.1373/clinchem.2010.150631.
98. Baxter RC. Inhibition of the insulin-like growth factor (IGF)-IGF-binding protein interaction. Horm Res 2001;55 Suppl 2:68-72. DOI: 10.1159/000063479.
99. Frystyk J, Freda P, Clemmons DR. The current status of IGF-I assays--a 2009 update. Growth Horm IGF Res 2010;20(1):8-18. DOI: 10.1016/j.ghir.2009.09.004.
100. Varewijck AJ, Lamberts SW, van der Lely AJ, Neggers SJ, Hofland LJ, Janssen JA. The introduction of the IDS-iSYS total IGF-1 assay may have far-reaching consequences for diagnosis and treatment of GH deficiency. J Clin Endocrinol Metab 2015;100(1):309-16. DOI: 10.1210/jc.2014-2558.
101. Bidlingmaier M, Friedrich N, Emeny RT, et al. Reference intervals for insulin-like growth factor-1 (igf-i) from birth to senescence: results from a multicenter study using a new automated chemiluminescence IGF-I immunoassay conforming to recent international recommendations. J Clin Endocrinol Metab 2014;99(5):1712-21. DOI: 10.1210/jc.2013-3059.
102. Chanson P, Arnoux A, Mavromati M, et al. Reference Values for IGF-I Serum Concentrations: Comparison of Six Immunoassays. J Clin Endocrinol Metab 2016;101(9):3450-8. DOI: 10.1210/jc.2016-1257.
103. Mavromati M, Kuhn E, Agostini H, et al. Classification of Patients With GH Disorders May Vary According to the IGF-I Assay. J Clin Endocrinol Metab 2017;102(8):2844-2852. DOI: 10.1210/jc.2017-00202.
104. Bidlingmaier M, Valcour A, Schilbach K, et al. Differences in the Distribution of IGF-I Concentrations Between European and US Populations. J Endocr Soc 2022;6(7):bvac081. DOI: 10.1210/jendso/bvac081.

105. Maione L, Albrici C, Grunenwald S, et al. IGF-I Variability Over Repeated Measures in Patients With Acromegaly Under Long-Acting Somatostatin Receptor Ligands. J Clin Endocrinol Metab 2022;107(9):e3644-e3653. DOI: 10.1210/clinem/dgac385.
106. Bystrom C, Sheng S, Zhang K, Caulfield M, Clarke NJ, Reitz R. Clinical utility of insulin-like growth factor 1 and 2; determination by high resolution mass spectrometry. PLoS One 2012;7(9):e43457. DOI: 10.1371/journal.pone.0043457.
107. Cox HD, Lopes F, Woldemariam GA, et al. Interlaboratory agreement of insulin-like growth factor 1 concentrations measured by mass spectrometry. Clin Chem 2014;60(3):541-8. DOI: 10.1373/clinchem.2013.208538.
108. Moncrieffe D, Cox HD, Carletta S, et al. Inter-Laboratory Agreement of Insulin-like Growth Factor 1 Concentrations Measured Intact by Mass Spectrometry. Clin Chem 2020;66(4):579-586. DOI: 10.1093/clinchem/hvaa043.
109. Mazzaferri EL. Management of a solitary thyroid nodule. N Engl J Med 1993;328(8):553-9. DOI: 10.1056/NEJM199302253280807.
110. Reiners C, Wegscheider K, Schicha H, et al. Prevalence of thyroid disorders in the working population of Germany: ultrasonography screening in 96,278 unselected employees. Thyroid 2004;14(11):926-32. DOI: 10.1089/thy.2004.14.926.
111. Dean DS, Gharib H. Epidemiology of thyroid nodules. Best Pract Res Clin Endocrinol Metab 2008;22(6):901-11. DOI: 10.1016/j.beem.2008.09.019.
112. Guth S, Theune U, Aberle J, Galach A, Bamberger CM. Very high prevalence of thyroid nodules detected by high frequency (13 MHz) ultrasound examination. Eur J Clin Invest 2009;39(8):699-706. DOI: 10.1111/j.1365-2362.2009.02162.x.
113. Tunbridge WM, Evered DC, Hall R, et al. The spectrum of thyroid disease in a community: the Whickham survey. Clin Endocrinol (Oxf) 1977;7(6):481-93. DOI: 10.1111/j.1365-2265.1977.tb01340.x.
114. Mu C, Ming X, Tian Y, et al. Mapping global epidemiology of thyroid nodules among general population: A systematic review and meta-analysis. Front Oncol 2022;12:1029926. DOI: 10.3389/fonc.2022.1029926.
115. Ross DS. Nonpalpable thyroid nodules--managing an epidemic. J Clin Endocrinol Metab 2002;87(5):1938-40. DOI: 10.1210/jcem.87.5.8552.
116. Sharbidre KG, Lockhart ME, Tessler FN. Incidental Thyroid Nodules on Imaging: Relevance and Management. Radiol Clin North Am 2021;59(4):525-533. DOI: 10.1016/j.rcl.2021.03.004.
117. Grani G, Sponziello M, Filetti S, Durante C. Thyroid nodules: diagnosis and management. Nat Rev Endocrinol 2024;20(12):715-728. DOI: 10.1038/s41574-024-01025-4.
118. Russ G, Leboulleux S, Leenhardt L, Hegedus L. Thyroid incidentalomas: epidemiology, risk stratification with ultrasound and workup. Eur Thyroid J 2014;3(3):154-63. DOI: 10.1159/000365289.
119. Durante C, Costante G, Lucisano G, et al. The natural history of benign thyroid nodules. JAMA 2015;313(9):926-35. DOI: 10.1001/jama.2015.0956.
120. Shetty SK, Maher MM, Hahn PF, Halpern EF, Aquino SL. Significance of incidental thyroid lesions detected on CT: correlation among CT, sonography, and pathology. AJR Am J Roentgenol 2006;187(5):1349-56. DOI: 10.2214/AJR.05.0468.
121. Frates MC, Benson CB, Doubilet PM, et al. Prevalence and distribution of carcinoma in patients with solitary and multiple thyroid nodules on sonography. J Clin Endocrinol Metab 2006;91(9):3411-7. DOI: 10.1210/jc.2006-0690.

122. Soelberg KK, Bonnema SJ, Brix TH, Hegedus L. Risk of malignancy in thyroid incidentalomas detected by 18F-fluorodeoxyglucose positron emission tomography: a systematic review. Thyroid 2012;22(9):918-25. DOI: 10.1089/thy.2012.0005.
123. Baloch ZW, Asa SL, Barletta JA, et al. Overview of the 2022 WHO Classification of Thyroid Neoplasms. Endocr Pathol 2022;33(1):27-63. DOI: 10.1007/s12022-022-09707-3.
124. Li M, Dal Maso L, Vaccarella S. Global trends in thyroid cancer incidence and the impact of overdiagnosis. Lancet Diabetes Endocrinol 2020;8(6):468-470. DOI: 10.1016/S2213-8587(20)30115-7.
125. Leboulleux S, Tuttle RM, Pacini F, Schlumberger M. Papillary thyroid microcarcinoma: time to shift from surgery to active surveillance? Lancet Diabetes Endocrinol 2016;4(11):933-942. DOI: 10.1016/S2213-8587(16)30180-2.
126. Mehanna H, Al-Maqbili T, Carter B, et al. Differences in the recurrence and mortality outcomes rates of incidental and nonincidental papillary thyroid microcarcinoma: a systematic review and meta-analysis of 21 329 person-years of follow-up. J Clin Endocrinol Metab 2014;99(8):2834-43. DOI: 10.1210/jc.2013-2118.
127. Durante C, Hegedus L, Czarniecka A, et al. 2023 European Thyroid Association Clinical Practice Guidelines for thyroid nodule management. Eur Thyroid J 2023;12(5). DOI: 10.1530/ETJ-23-0067.
128. Borson-Chazot F, Borget I, Mathonnet M, Leenhardt L. SFE-AFCE-SFMN 2022 consensus on the management of thyroid nodules: Epidemiology and challenges in the management of thyroid nodules. Ann Endocrinol (Paris) 2022;83(6):378-379. DOI: 10.1016/j.ando.2022.10.003.
129. Borson-Chazot F, Buffet C, Decaussin-Petrucci M, et al. SFE-AFCE-SFMN 2022 consensus on the management of thyroid nodules: Synthesis and algorithms. Ann Endocrinol (Paris) 2022;83(6):440-453. DOI: 10.1016/j.ando.2022.11.001.
130. Grani G, Sponziello M, Pecce V, Ramundo V, Durante C. Contemporary Thyroid Nodule Evaluation and Management. J Clin Endocrinol Metab 2020;105(9):2869-83. DOI: 10.1210/clinem/dgaa322.
131. Tessler FN, Middleton WD, Grant EG, et al. ACR Thyroid Imaging, Reporting and Data System (TI-RADS): White Paper of the ACR TI-RADS Committee. J Am Coll Radiol 2017;14(5):587-595. DOI: 10.1016/j.jacr.2017.01.046.
132. Russ G, Bonnema SJ, Erdogan MF, Durante C, Ngu R, Leenhardt L. European Thyroid Association Guidelines for Ultrasound Malignancy Risk Stratification of Thyroid Nodules in Adults: The EU-TIRADS. Eur Thyroid J 2017;6(5):225-237. DOI: 10.1159/000478927.
133. Haugen BR, Alexander EK, Bible KC, et al. 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: The American Thyroid Association Guidelines Task Force on Thyroid Nodules and Differentiated Thyroid Cancer. Thyroid 2016;26(1):1-133. DOI: 10.1089/thy.2015.0020.
134. Ha EJ, Chung SR, Na DG, et al. 2021 Korean Thyroid Imaging Reporting and Data System and Imaging-Based Management of Thyroid Nodules: Korean Society of Thyroid Radiology Consensus Statement and Recommendations. Korean J Radiol 2021;22(12):2094-2123. DOI: 10.3348/kjr.2021.0713.
135. Durante C, Hegedus L, Na DG, et al. International Expert Consensus on US Lexicon for Thyroid Nodules. Radiology 2023;309(1):e231481. DOI: 10.1148/radiol.231481.

136. Ali SZ, Baloch ZW, Cochand-Priollet B, Schmitt FC, Vielh P, VanderLaan PA. The 2023 Bethesda System for reporting thyroid cytopathology. J Am Soc Cytopathol 2023;12(5):319-325. DOI: 10.1016/j.jasc.2023.05.005.
137. Castellana M, Castellana C, Treglia G, et al. Performance of Five Ultrasound Risk Stratification Systems in Selecting Thyroid Nodules for FNA. J Clin Endocrinol Metab 2020;105(5). DOI: 10.1210/clinem/dgz170.
138. Grani G, Lamartina L, Ascoli V, et al. Reducing the Number of Unnecessary Thyroid Biopsies While Improving Diagnostic Accuracy: Toward the "Right" TIRADS. J Clin Endocrinol Metab 2019;104(1):95-102. DOI: 10.1210/jc.2018-01674.
139. Persichetti A, Di Stasio E, Coccaro C, et al. Inter- and Intraobserver Agreement in the Assessment of Thyroid Nodule Ultrasound Features and Classification Systems: A Blinded Multicenter Study. Thyroid 2020;30(2):237-242. DOI: 10.1089/thy.2019.0360.
140. Russ G, Trimboli P, Buffet C. The New Era of TIRADSs to Stratify the Risk of Malignancy of Thyroid Nodules: Strengths, Weaknesses and Pitfalls. Cancers (Basel) 2021;13(17). DOI: 10.3390/cancers13174316.
141. Tessler FN, Thomas J. Artificial Intelligence for Evaluation of Thyroid Nodules: A Primer. Thyroid 2023;33(2):150-158. DOI: 10.1089/thy.2022.0560.
142. Toro-Tobon D, Loor-Torres R, Duran M, et al. Artificial Intelligence in Thyroidology: A Narrative Review of the Current Applications, Associated Challenges, and Future Directions. Thyroid 2023;33(8):903-917. DOI: 10.1089/thy.2023.0132.
143. Li Y, Liu Y, Xiao J, et al. Clinical value of artificial intelligence in thyroid ultrasound: a prospective study from the real world. Eur Radiol 2023;33(7):4513-4523. DOI: 10.1007/s00330-022-09378-y.
144. Potipimpanon P, Charakorn N, Hirunwiwatkul P. A comparison of artificial intelligence versus radiologists in the diagnosis of thyroid nodules using ultrasonography: a systematic review and meta-analysis. Eur Arch Otorhinolaryngol 2022;279(11):5363-5373. DOI: 10.1007/s00405-022-07436-1.
145. Mavromati M, Saiji E, Demarchi MS, et al. Unnecessary thyroid surgery rate for suspicious nodule in the absence of molecular testing. Eur Thyroid J 2023;12(6). DOI: 10.1530/ETJ-23-0114.
146. Chiosea S, Hodak SP, Yip L, et al. Molecular Profiling of 50 734 Bethesda III-VI Thyroid Nodules by ThyroSeq v3: Implications for Personalized Management. J Clin Endocrinol Metab 2023;108(11):2999-3008. DOI: 10.1210/clinem/dgad220.
147. Livhits MJ, Zhu CY, Kuo EJ, et al. Effectiveness of Molecular Testing Techniques for Diagnosis of Indeterminate Thyroid Nodules: A Randomized Clinical Trial. JAMA Oncol 2021;7(1):70-77. DOI: 10.1001/jamaoncol.2020.5935.
148. Vuong HG, Nguyen TPX, Hassell LA, Jung CK. Diagnostic performances of the Afirma Gene Sequencing Classifier in comparison with the Gene Expression Classifier: A meta-analysis. Cancer Cytopathol 2021;129(3):182-189. DOI: 10.1002/cncy.22332.